Traitement du Signal
Vol. 42, No. 4, August, 2025, pp. 2205-2214
Journal homepage: http://iieta.org/journals/ts

GuitarNeXt: An Advanced Convolutional Neural Network Architecture for Music Genre Classification

Peijin Du, Zhenhua Zhou*, Fengquan Li

Xi'an International University, Xi'an 710077, China

Corresponding Author Email: zhou20218108@163.com

https://doi.org/10.18280/ts.420431

Received: 8 January 2025
Revised: 29 May 2025
Accepted: 18 June 2025
Available online: 14 August 2025

Keywords:
GuitarNeXt, musical genre classification, deep learning

ABSTRACT

Music genre classification is a challenging task that has been extensively addressed using various deep learning methods. Recently, convolutional neural networks (CNNs) have shown significant promise in this domain. This paper introduces GuitarNeXt, a novel CNN architecture designed specifically for music genre classification. Our approach utilizes a publicly available dataset containing audio recordings and spectral images across multiple genres to evaluate the performance of GuitarNeXt. The architecture of GuitarNeXt includes four primary layers: a stem layer employing patchify convolution to produce initial tensors, a main GuitarNeXt layer that integrates a hybrid attention mechanism with depth concatenation and scaling convolutions, a downsampling layer combining average and maximum pooling with depth concatenation, and an output layer that applies global pooling to produce a feature map for classification. Experimental results demonstrate that GuitarNeXt achieves a classification accuracy of 96.40% and a precision of 96.59% on the test set, highlighting its effectiveness and potential as a robust tool for automated music genre classification. This innovative model not only advances the field of music analysis but also sets a new benchmark for subsequent research in the area.

1. INTRODUCTION

In the field of Music Information Retrieval (MIR), various applications such as music genre classification, music streaming services, automatic playlist generation and music recommendation systems are important [1, 2]. With the rapid expansion of music archives digitally, the accurate classification of music genres has become increasingly important [3]. Features manually extracted from audio signals form the basis of traditional classification methods [4]. However, these methods limit the classification accuracy as they cannot adequately capture the hierarchical and complex patterns of music [5]. Recent rapid advances in deep learning have provided better results in classifying music genres using features learned directly from raw audio signals or their spectrogram representations [6]. Convolutional Neural Networks (CNN) have been widely used in this field due to their ability to identify spatially correlated features from spectrogram images [7]. However, currently used CNN-based approaches often use traditional architectures and there are no next-generation CNN architectures specifically designed for music genre classification [8]. Deep learning efforts for music genre classification have focused on adapting common CNN architectures developed for image recognition [9]. Traditional CNNs have fundamental limitations such as local receptive fields and the inability to model long-range dependencies. Although these models are successful [10]. Since the temporal structures and harmonic patterns of music span wide frequency ranges, music analysis is very challenging [11]. It has recently emerged that advanced approaches such as self-attention mechanisms and hybrid deep learning architectures can overcome these limitations [12]. However, the field of music classification has not sufficiently explored these innovative approaches. Therefore, there is a need for a creative CNN model specifically designed for classifying music genres [9, 13].

This paper proposes a new CNN model known as GuitarNeXt. GuitarNeXt uses a hybrid attention block that can capture both local and international dependencies using convolution-based attention mechanisms and pooling methods. These methods are inspired by ConvertNeXt and Transformer-based architectures. Unlike traditional CNNs, the GuitarNeXt model includes a Patchify-based input layer, depth concatenation, and a hybrid downsampling strategy that includes maximum mean pooling. These design choices help the model extract rich and diverse features from spectrogram images and improve computational efficiency.

The vast majority of existing models for deep learning-based music classification are based on traditional architectures and are specifically designed for spectrogram-based classification [14]. Audio and music are structures with complex features, and capturing these complex features is challenging, even though it has been attempted to be solved with the use of state-of-the-art architectures [15, 16]. Traditional CNN architectures are not suitable for real-time applications due to the computational cost and accuracy complexities encountered during classification [17]. GuitarNeXt aims to overcome these limitations with its hybrid

attention mechanism and optimized subsampling strategy to create a powerful and high performance framework for music genre classification. Considering the innovative nature of Guitar NeXt, in addition to its high classification accuracy, it is also a preliminary work for applications such as music information retrieval and deep learning-based audio analysis.

The developed model can be successfully used in various music domains such as instrument recognition, music genre analysis, and rhythm classification. GuitarNeXt has a powerful architecture despite its small size. Therefore, it can increase its usability in real-time and efficient audio classification tasks. In response to the increasing demand in this field, it enables the development of scalable and customizable deep learning models for audio classification solutions. This study used publicly available music genre classification data and trainings to evaluate the performance of the GuitarNeXt architecture. The results show that the proposed technique achieves high accuracy rates containing approximately 4.7 million parameters. In addition, the proposed technique achieved better results in terms of classification accuracy compared to other CNN architectures. When the obtained results are examined, it is possible to solve complex audio problems such as music classification with high accuracy with the combination of hybrid attention mechanism and innovative subsampling methods.

1.1 Literature review

Music genre classification is an important research topic for audio processing and artificial intelligence applications [18, 19]. While advanced methods rely on manually designed features, deep learning models such as CNN and Transformer have improved accuracy by better analyzing spectral and temporal features [20]. In the literature, unimodal methods combine audio, visual and text data. Lighter neural network architectures are preferred to increase efficiency in resource-constrained systems [21]. This literature review provides an overview of recent advances and challenges in music genre classification.

Adamczyk et al. [22] investigated machine learning methods in automatic music generation and classification. Deep learning models include Generative Adversarial Networks (GAN), Long Short Term Memory (LSTM) and Convolutional Neural Networks (CNN). Transformer-based models aim to analyze and synthesize music. The research examined the ability of music genre classification models to accurately classify musical tracks using the Free Music Archive (FMA) and GTZAN datasets. The transformer-based model used in the study achieved only 30% accuracy on the GTZAN dataset, while the CNN-based classifier achieved 86% accuracy. Although computational resources are limited, Transformer architectures can be used in music analysis and production. Additionally, future hardware optimizations are said to have the potential to improve performance.

Levin and Singer [23] developed a study in which they characterized the Graph-Based Automatic Feature Selection method as multiclass. Jeffries-Matusita (JM) distance and t-distributed Stochastic Neighbor Embedding (t-SNE) techniques were used in the study to improve classification accuracy. With these procedures, it is aimed to achieve high accuracy rates with less data. In the study, attribute clustering results were evaluated using the MSS index and it was aimed to create a subset of attributes with the most prominent features without the need for predefined selection parameters. Experiments were conducted with various datasets, including the music genre classification dataset, and the proposed method maintained the classification performance despite feature reduction and reduced the computational cost by 15% to 70%. These results were obtained by using a feature km of between 7% and 30%. It has been shown that graph-based feature selection can be used to improve the efficiency of deep learning models in complex tasks such as music classification.

Ba et al. [24] conducted a study on the classification of music genres using deep learning techniques. In their research, various deep learning architectures such as capsule neural networks (CSN), convolutional neural networks (CNN), Long Short Term Memory (LSTM) networks and gated recurrent units (GRU) were used. They aimed to provide an innovative approach by combining the use of these architectures with mel-spectogram-based feature extraction. The study was carried out on the GTZAN dataset, which is widely used in this field, and they performed preprocessing steps for data augmentation on the data. As a result of the experiments, it was observed that the CSN model achieved 99.91% accuracy.

Elbir and Aydin [25] proposed a system that can perform signal processing and music classification using deep learning methods. The proposed method, named MusicResNet, was developed to extract acoustic features of both audio signals and Mel spectograms. GTZAN dataset containing 10 different music genres was used in the study. As a result of the training of the developed method with the obtained data, an accuracy of 81.8% was achieved. In addition, a hybrid method was proposed by combining the proposed method with Support Vector Machine (SVM) and the accuracy rate was increased to 97.6%. The results of the experiments show that the proposed hybrid method is very effective for music genre classification.

Wu et al. [26] proposed a method for underwater acoustic voiceprint recognition. Their model, called Echo Lite Voice Fusion Network (ELVFN), combines deep separable convolution and self-attention mechanisms to provide a new approach. This approach aims to reduce computational complexity and increase accuracy. As a result of the experiments, the proposed method achieved 94.66% test success. Compared to traditional models such as CAM++, TDNN and ECAPA-TDNN, the method was found to work more accurately and efficiently.

In a study on multimodal music genre classification, Oguike and Primus [27] developed a classification system for Sotho-Tswana music videos. The study developed a system that can predict genre using audio, lyrics and visual data. Artificial Neural Networks (ANN), visual mode VGG16 and text mode BERT were used. The separate predictions of each mode were combined using a late fusion technique to obtain the final species prediction. The multimodal model using late fusion achieved 94.8% accuracy, while the audio, visual and text models achieved 85.2%, 90.4% and 78.6% accuracy respectively. Multimodal music genre classification is better than unimodal methods.

Wang et al. [28] proposed an intelligent music classification method that combines deep learning-based classification and feature extraction. In the study, VGG-16 Net was used to classify Bi-directional Long Short-Term Memory (BiLSTM) networks. MSD-I, GTZAN and ISMIR2004 type datasets were used for the research. Outperforming traditional models (SVM and KNN), the BiLSTM-VGG-16 net model achieved 97% accuracy in MSD-I, 97.8% in GTZAN and 96.5% in ISMIR2004 [16].

1.2 Literature gaps

The identified literature gaps are:

•In the literature, there are various automatic biomedical image classification models [29-31] but there are a few music classification models.

•The most researchers have utilized the well-known deep learning architectures in researchers to attain high classification performances [32-34]. The presented most of the custom CNN models are VGG-like models [35, 36]. Thus, there is a stagnation in the new-generation deep learning model proposing.

1.3 Motivation and study outline

The main motivation of this work is to recommend an innovative music-inspired CNN model and investigate this CNN's classification ability using music genre classification. In this CNN architecture, to achieve high classification performance, we have introduced a new attention structure.

To fill the first identified gap in Section 1.2, we have introduced a new music-based deep learning model. Therefore, we have used a publicly available music genre classification dataset and reported the classification results of the proposed CNN on this dataset.

To address the second gap, GuitarNeXt, a new-generation CNN model, has been introduced, and its results have been evaluated on the music genre dataset. GuitarNeXt has an original structure. Therefore, we have introduced an innovative CNN in this research.

To develop GuitarNeXt, we were inspired by ConvNeXt and transformers. The ConvNeXt structure is an effective CNN architecture. Moreover, transformer models have achieved high performance due to their attention blocks. Therefore, the proposed GuitarNeXt is a fully convolutional deep learning model like ConvNeXt, but we have introduced an attention block using convolutions and pooling (similar to PoolFormer). The proposed GuitarNeXt consists of four main phases: (i) stem, (ii) main, (iii) downsampling, and (iv) output. The stem block used is similar to that in ConvNeXt, while the other three phases (main, downsampling, and output) have unique features.

1.4 Innovation and contributions

Innovation:

•A new-generation attention block has been presented in this research.

•We have presented hybrid pooling-based approximation and this approximation has been utilized in the downsampling and flatten layers of the GuitarNeXt.

•By utilizing the presented attention block, GuitarNeXt layer has been created.

•Deploying GuitarNeXt layer as main block, the presented GuitarNeXt CNN has been presented.

Contributions:

•CNN research area is one of the most important research areas for deep learning especially computer vision. However, transformers have eclipsed CNNs in computer vision since transformers can attain higher classification performances. Although, CNN research area is a young research area and many advancements can be discovered in this research area. To contribute the CNN research area, the GuitarNeXt has been presented.

•In this research, a publicly available music genre classification dataset has been utilized to contribute musical deep learning research area.

2. THE USED DATASET

In the study, the 10-class GTZAN- Music Genre Classification dataset shared by Andrada Olteanu on the Kaggle platform was used [37]. GTZAN dataset is one of the most widely used open source datasets for music genre recognition. This dataset is divided into classes consisting of 10 different music genres: blues, classical, country disco, hiphop, jazz, metal, pop, reggae and rock. The data belonging to the 10 classes is divided into 100 audio files in each class with a balanced distribution. Each audio file in the dataset is 30 seconds long and has been brought together from various conditions such as CD, radio recordings, microphone recordings in order to provide sounds in each and different conditions. The data in the dataset is divided into audio and spectogram data; spectogram images are also used in the study.

3. PROPOSED GUITARNEXT

In this research, we have presented an innovative deep learning model, GuitarNeXt. Our main objective is to discover new attributes of CNNs. Therefore, we have introduced an attention mechanism-based fully convolutional deep learning architecture. To create this CNN architecture, we were inspired by ConvNeXt (for CNN architecture design), transformers (as their attention layers are highly effective, and ConvNeXt itself was inspired by the Swin-Transformer), and PoolFormer (where we employed pooling-like convolutions to create the attention block). This CNN structure consists of four main phases, which are:

•Patchify-based stem,

•Main/GuitarNeXt phase,

•Average and maximum pooling-based downsampling,

•Global average and maximum pooling-based output.

To give a details about to presented GuitarNeXt, we have demonstrated schematic illustration of this CNN in Figure 1.

Per Figure 1, our model has four main phases and details of these phases are explained below.

3.1 Stem phase

The stem phase is the first phase of the proposed CNN (GuitarNeXt). Patchify convolution and batch normalization have been used. In this phase, an image of size 224×224×3 is transformed into a tensor of size 56×56×96. The mathematical definition of this stem phase is provided below.

$$T_1 = BN\left(C^{96}_{4\times4,Stride:4}(Im)\right) \quad (1)$$

where, T: the created tensor, $BN(.)$: Batch Normalization function, $C(.)$: the patchify convolution and Im: the utilized image with a size of 224×224×3.

3.2 GuitarNeXt phase

The main phase of the proposed CNN is GuitarNeXt. The introduced GuitarNeXt is an attention-based layer. In this phase, convolution, maximum pooling, average pooling, depth

concatenation, and GELU activation are utilized. We have applied this phase four times to generate a feature map, and its mathematical definition is given below.

$$T_n = GC^F_{7\times7}(T_{n-1}) \tag{2}$$

where, $GC(.)$: grouped convolution and F: the number of the filters. Above, the first tensor of the proposed main stage is shown, where we have used a 7×7 convolution similar to ConvNeXt. Below, we present pooling and convolution-based attention.

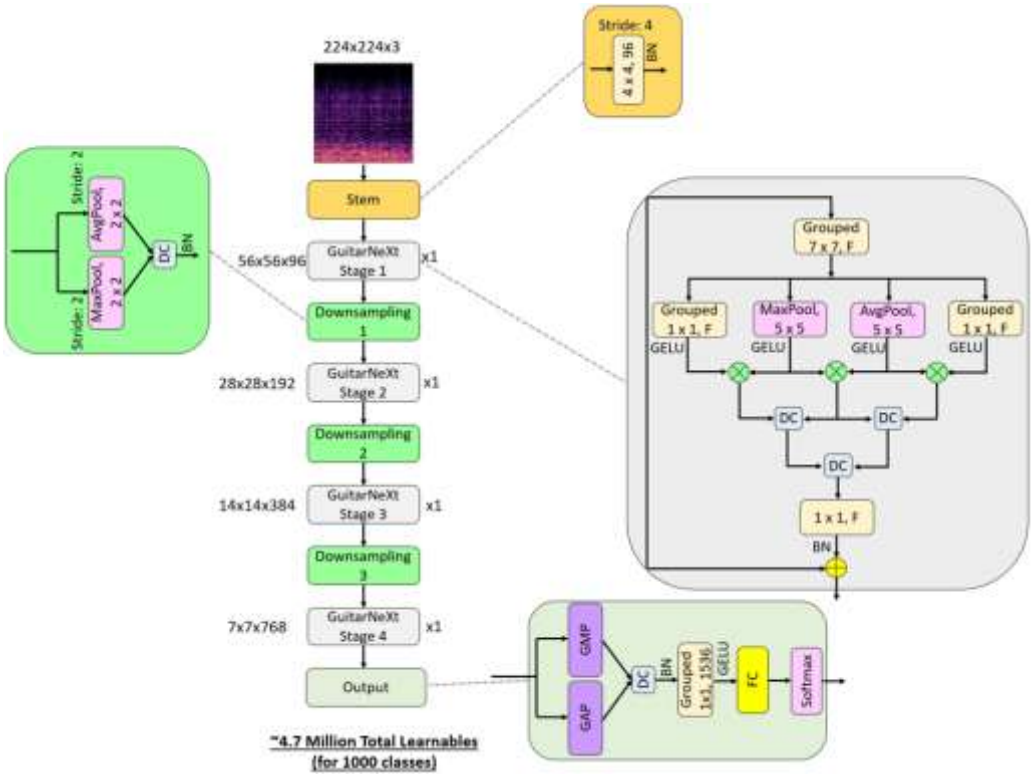


Figure 1. The graphical outline of the recommended GuitarNeXt
Note: BN - Batch Normalization, DC - Depth Concatenation, GELU - Gaussian Error Linear Unit, F - the number of filters, GMP - Global Maximum Pooling, GAP - Global Pooling, FC - Fully Connected

$$T^1_{n+1} = GELU\left(GC^F_{1\times1}(T_n)\right)\otimes GELU\left(MP_{5\times5}(T_n)\right) \tag{3}$$

$$T^2_{n+1} = GELU\left(AP_{5\times5}(T_n)\right)\otimes GELU\left(MP_{5\times5}(T_n)\right) \tag{4}$$

$$T^3_{n+1} = GELU\left(GC^F_{1\times1}(T_n)\right)\otimes GELU\left(AP_{5\times5}(T_n)\right) \tag{5}$$

where, $GELU(.)$: the GELU activation function, $AP(.)$: the average pooling, $MP(.)$: maximum pooling and $\otimes$: multiplaction. Here, the stride value of this pooling function is set to one. To create an inverted bottleneck, we have used depth concatenation to increase the number of filters from F to 4F. The depth concatenation step is mathematically defined below.

$$T^1_{n+2} = DC(T^1_{n+1}, T^2_{n+1}) \tag{6}$$

$$T^2_{n+2} = DC(T^1_{n+1}, T^3_{n+1}) \tag{7}$$

$$T_{n+3} = DC(T^1_{n+2}, T^2_{n+2}) \tag{8}$$

where, $DC(.)$: the depth concatenation function.

The last step is the scaling and shortcut step. We have used this step to address the vanishing gradient problem. The equations for this step are given below.

$$T_{n+4} = BN\left(C^F_{1\times1}(T_{n+3})\right) \tag{9}$$

$$T_{n+5} = T_{n+4} + T_{n-1} \tag{10}$$

where, T_{n+5} is the output of the GuitarNeXt section.

3.3 Downsampling phase

In this research, we have introduced a new hybrid downsampling stage, utilizing maximum pooling, average pooling, and depth concatenation to address the routing problem. This phase is mathematically defined below.

$$T_d = BN\left(DC\left(MP^F_{2\times2,Stride:2}(T_{d-1}), DC\left(AP^F_{2\times2,Stride:2}(T_{d-1})\right)\right)\right) \tag{11}$$

In this phase, the tensor's width and height are halved, while its depth is doubled.

3.4 Output phase

The classification output of the proposed GuitarNeXt is generated in this phase. This phase includes a hybrid flattening step. Therefore, global average pooling (GAP) and global maximum pooling (GMP) are used together. Afterward, a grouped convolution and GELU activation are applied as post-processing to the generated feature map. The processed feature map is then classified using fully connected and softmax layers. The mathematical definitions of the output phase are given below.

$$flat = DC\left(GAP(T_{last}), GMP(T_{last})\right) \tag{12}$$

By utilizing both GAP and GMP functions together, we have obtained the flatten feature vector. Herein, $flat$: the flatten feature vector, $GAP(.)$: Global Average Pooling, $GMP(.)$: Global Maximum Pooling and T_{last}: the last tensor created.

To apply post-processing to the flattened feature vector, we have used grouped convolution and GELU activation.

$$fv = GELU\left(GC_{2\times2}^{1536}(flat)\right) \tag{13}$$

where, fv: the created last feature vector. In order to obtain the classification outcome from this feature vector, we have applied to fully connected (to determine the number of the features) and softmax (to generate the classification outcome) operators.

$$out = SM\big(FC(fv)\big) \tag{14}$$

where, out: the classification outcome, $SM(.)$: softmax and $FC(.)$: fully connected.

By utilizing this step, we have presented the lightweight version of the GuitarNeXt. Moreover, the steps the presented GuitarNeXt CNN have been given in subsequent section.

3.5 Overview of the presented GuitarNeXt

To more clearly explain the proposed GuitarNeXt, we have outlined the steps of this lightweight CNN in this section. Each phase of the CNN is defined as a mathematical function. These steps are:

Step 1: Create the first tensor by applying the stem function, resulting in a tensor of size 56×56×96.

$$T_1 = stem(Im) \tag{15}$$

where, $stem(.)$: the function defines the stem phase and T_1: the first tensor.

Step 2: Apply the first GuitarNeXt phase by utilizing the first tensor (T_1) as input.

$$T_2 = GN(T_1) \tag{16}$$

where, $GN(.)$: the function defines the GuitarNeXt/main phase and T_2: the second tensor with a size of 56×56×96.

Step 3: Use downsampling to create a new tensor with a size of 28×28×192.

$$T_3 = DS(T_2) \tag{17}$$

where, $DS(.)$: the function defines the downsampling phase and T_3: the first down sampled tensor with a size of 28×28×192.

Step 4: Utilize the presented main/GuitarNeXt function to generate features from the down sampled tensor.

$$T_4 = GN(T_3) \tag{18}$$

The size of the T_4 is 28×28×192

Step 5: Employ downsampling to create a tensor with a size of 14×14×384.

$$T_5 = DS(T_4) \tag{19}$$

T_5: the second down sampled tensor with a size of 14×14×384.

Step 6: Deploy the GuitarNeXt function to extract feature map from the down sampled tensor.

$$T_6 = GN(T_5) \tag{20}$$

The size of the T_6 is 14×14×384.

Step 7: Employ the third downsampling to create a tensor with a size of 7×7×768.

$$T_7 = DS(T_6) \tag{21}$$

T_7: the third down sampled tensor with a size of 7×7×768.

Step 8: Apply the GuitarNeXt function to the third down sampled tensor.

$$T_8 = GN(T_7) \tag{22}$$

The size of the T_8 is 7×7×768 and T_8 is defined as the last tensor and it is utilized as input of the output/classification phase.

Step 9: Use output function (it is defined in Section 3.4) to obtain classification outcomes.

$$out = OT(T_8) \tag{23}$$

where, out: the classification outcome, $OT(.)$: the function defined output phase/stage.

According to Steps 1-9, the common mathematical definition of the introduced GuitarNeXt is explained below.

$$\begin{gathered} R: \{1,1,1,1,1\}, F: (96,192,384,768,1536), \\ WH: (56^2, 28^2, 14^2, 7^2, 1^2) \end{gathered} \tag{24}$$

where, R: the number of repetitions, F: the size of the filters and WH: the width and height of the generated tensors. "In this GuitarNeXt, each phase is repeated once. By deploying the stem, GuitarNeXt, and downsampling phases, we have generated a feature map of size 7×7×768. In the flattening layer, by utilizing GAP, GMP, and depth concatenation, the number of filters is increased from 768 to 1536.

4. EXPERIMENTAL RESULTS

To evaluate GuitarNeXt's classification ability, this CNN has been applied to the music genre classification dataset. Since this is a computer vision task, we have used power spectrogram images of the corresponding WAV signals. In this research, we have presented a lightweight version of the proposed CNN. As a result, the introduced version of GuitarNeXt has approximately 4.7 million learnable parameters for 1,000 classes. The transition table of the proposed model is given in Table 1.

The proposed GuitarNeXt CNN was developed using MATLAB version 2023a Deep Network Designer. This tool is a scratch-based tool, and we used blocks and connections to construct the proposed GuitarNeXt CNN. To implement this CNN model in MATLAB Deep Network Designer, we used a personal computer (PC) equipped with an NVIDIA GeForce 4090 graphics processing unit (GPU).

In the training phase, we used the following parameters, which are listed in Table 2.

Table 1. The transitions of the presented GuitarNeXt

Step	Output Size	Process
Stem	56×56	4×4, 96, stride 4
GuitarNeXt 1	56×56	$\begin{bmatrix} GC7\times 7{,}96 \\ GC1\times 1{,}96, MP5\times 5{,}96, AP5\times 5{,}96, GC1\times 1{,}96 \\ 1\times 1{,}96 \end{bmatrix}\times 1$
Downsampling 1	28×28	2×2 maximum pooling, 96, stride 2, 2×2 average pooling, 96, stride 2, depth concatenation
GuitarNeXt 2	28×28	$\begin{bmatrix} GC7\times 7{,}192 \\ GC1\times 1{,}192, MP5\times 5{,}192, AP5\times 5{,}192, GC1\times 1{,}192 \\ 1\times 1{,}192 \end{bmatrix}\times 1$
Downsampling 2	14×14	2×2 maximum pooling, 192, stride 2, 2×2 average pooling, 192, stride 2, depth concatenation
GuitarNeXt 3	14×14	$\begin{bmatrix} GC7\times 7{,}384 \\ GC1\times 1{,}384, MP5\times 5{,}384, AP5\times 5{,}384, GC1\times 1{,}384 \\ 1\times 1{,}384 \end{bmatrix}\times 1$
Downsampling 3	7×7	2×2 maximum pooling, 384, stride 2, 2×2 average pooling, 384, stride 2, depth concatenation
GuitarNeXt 4	7×7	$\begin{bmatrix} GC7\times 7{,}768 \\ GC1\times 1{,}768, MP5\times 5{,}768, AP5\times 5{,}768, GC1\times 1{,}768 \\ 1\times 1{,}768 \end{bmatrix}\times 1$
Output	NC	Flatten with GAP and GMP, depth concatenation $[GC1\times 1{,}1536]$ Fully connected, softmax.
The number of parameters		~4.7 million for 1000 classes.

** NC: the number of classes.

Table 2. The utilized training settings

Parameter	Value
Solver	Stochastic Gradient Descent Momentum (SGDM)
Initial learning rate	0.01
Mini batch size	128
Maximum epoch	30
L2 Regularization	0.001
Gradient threshold method	L2
Learning rate drop factor	0.1
Training and validation split ratio	80:20, randomize

Table 3. Data augmentation parameters

Parameter	Value
Image rotation	45,90,135,180,225,270
Noise addition	Salt and peppers with 0.05 and speckle

Moreover, we used augmentation to increase the number of images since the original dataset contained a limited number of images. The augmented images were used for training, while the original spectrogram images were utilized for testing. The data augmentation process is provided in Table 3.

By utilizing the above parameters, eight augmented images were created from each original image. In this context, the training, validation, and test split ratio for this research is computed as 70:17.5:12.5.

Moreover, we have used a three-layered training phase, and the steps of this training approach are as follows:

Step 1: Train the proposed GuitarNeXt on ImageNet1k and save the pretrained network as PGN1.

Step 2: Train PGN1 (created in Step 1) on the designated training dataset using augmentation with rotation. We applied rotations in the range of -25 to 20 degrees for augmentation. In this step, the newly trained pretrained network is saved as PGN2.

Step 3: Train the designated training dataset using PGN2 and obtain the final pretrained CNN.

By utilizing this strategy, we trained the music genre classification dataset, and the resulting training and validation curves are shown in Figure 2.

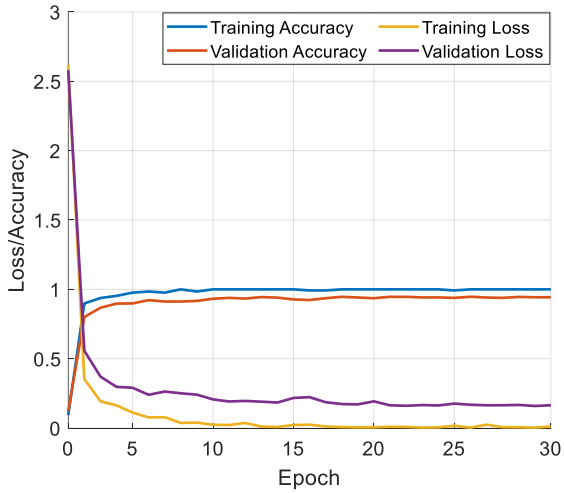


Figure 2. The training and validation curve of the recommended GuitarNeXt CNN

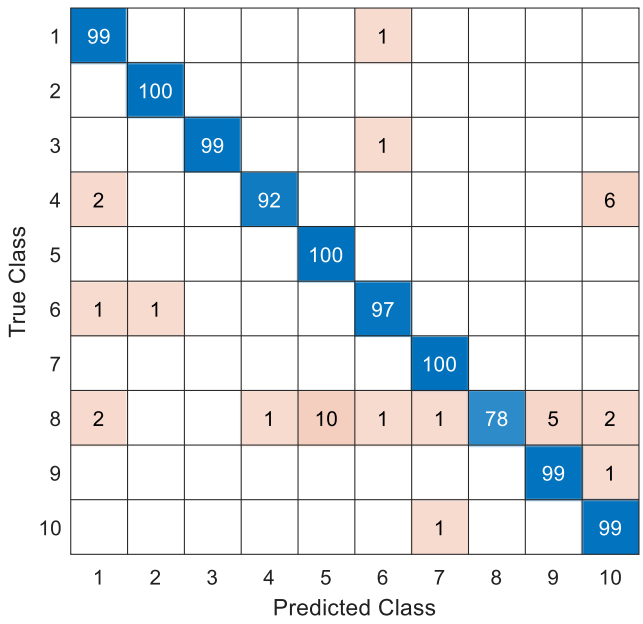


Figure 3. The computed test confusion matrix

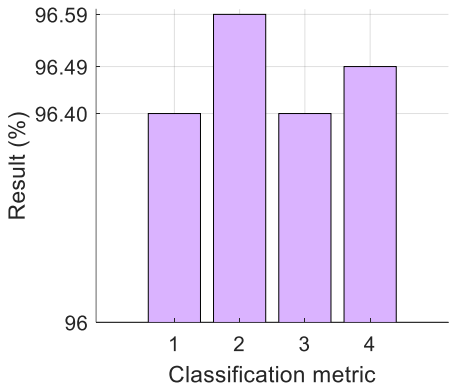
According to Figure 2, the GuitarNeXt CNN achieved a final training accuracy of 100%, a final training loss of 0.0118, a final validation accuracy of 95.93%, and a final validation loss of 0.1378. Using this trained GuitarNeXt (final pretrained version), the test confusion matrix is illustrated in Figure 3.

To evaluate the computed test classification output, the four commonly used classification performance evaluation metrics were utilized. These metrics are (i) classification accuracy, (ii) overall precision, (iii) unweighted average recall, and (iv) F1-score. The test classification results are presented in Table 4.

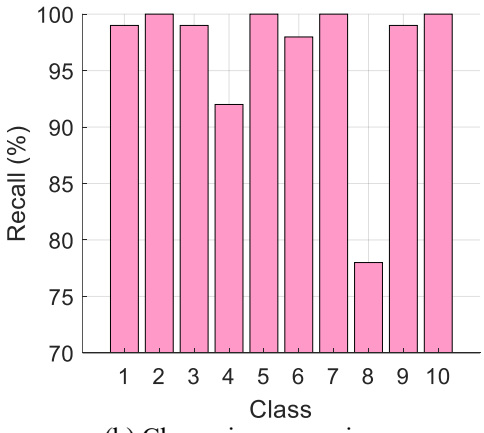
Table 4. The computed test results of the proposed GuitarNeXt on the utilized music genre dataset

No.	Classification metric	Result (%)
1	Accuracy	96.40
2	Overall Precision	96.59
3	Unweighted Average Recall	96.40
4	F1	96.49

These classification results and the class-wise accuracies (recall) are also depicted in Figure 4.



(a) Overall results, 1: Accuracy, 2: Precision, 3: Recall, 4: F1-score



(b) Class-wise accuracies.

Figure 4. Test results

As seen in Figure 4(a), the test precision was computed as 96.59%, while the test accuracy and unweighted average recall were both computed as 96.40%. The F1-score, which is the harmonic mean of recall and precision, was calculated as 96.49%.

Figure 4(b) highlights that the 2nd, 5th, 7th, and 10th classes achieved 100% (excellent) class-wise accuracy, while the worst-performing class was the 8th class, which yielded a recall of 78%.

5. DISCUSSIONS

In this research, we have presented a-new generation CNN architecture and this CNN is called GuitarNeXt since its shape is similar to guitar. We have presented an innovative hybrid attention in this GuitarNeXt model and this attention mechanism has been created deploying pooling and convolution operators together. The classification performance of the introduced GuitarNeXt was investigated on a music genre classification dataset.

The proposed GuitarNeXt architecture attained high classification performance on the music genre dataset and this CNN model (GuitarNeXt) yielded a test accuracy of 96.40%, a test overall precision of 96.59%, and a test F1-score of 96.49%. These results illustrate that the introduced hybrid attention mechanism—combining both pooling-based and convolution-based attention—effectively extracts the most informative features from the power spectrogram images. In particular, four classes (such as the 2nd, 5th, 7th, and 10th) reached a perfect 100% test recall.

One key finding is the role of the downsampling phase, and the presented downsampling phase uses a combination of maximum and average pooling along with depth concatenation. By utilizing this strategy, without using convolution, routing problem was solved. Additionally, the same strategy has been utilized in output phase. In this phase GAP and GMP have been employed in the flatten layer and by using these both functions (GAP and GMP) a reach feature vector has been computed. This strategy contributes to a richer representation by extracting both overall trends (average) and distinctive peaks (maximum) in the feature vectors. Also, the introduced GuitarNeXt achieve high classification performance with a relatively lightweight number of the learnable parameters and the GuitarNeXt has ~4.7 million learnable parameters.

Table 5. The comparative results for same dataset

Ref.	Study	Accuracy (%)
[22]	CNN	**86.00**
[38]	Multimodal CNN	92.20
[39]	Enhanced capsule neural network	90.40
[40]	Convolution temporal pooling network	75.00
[41]	Principal component analysis	77.41
[42]	Bidirectional LSTM	93.10
[43]	CNN-LSTM	87.50
[44]	CNN	74.10
[5]	Spectral and acoustic feature	81.50
[45]	CNN	93.40
[46]	Deep 1D CNN	82.03
[19]	CNN	92.70
[47]	Functional data analysis	84.50
[48]	AlexNet	95.38
[49]	CNN	75.20
[50]	Empirical mode decomposition	91.00
[51]	Neural network	84.00
[52]	Multilayer perceptron, CNN	95.50
[53]	CNN	81.00
Our study	GuitarNeXt	96.40

Another consideration involves the training strategy. While the three-stage approach (pre-training on ImageNet1k, then fine-tuning on augmented data, and finally refining with rotations) was successful. By using this strategy, our primary

objective is to apply distillation to get higher classification performance. In this aspect, this GuitarNeXt used the same methodology as DeepSeek.

In order to show the position of the presented model, we have compared the GuitarNeXt to other state-of-the-art (SOTA) models and these benchmark classification results have been listed in Table 5.

Table 5 openly highlighted that the introduced GuitarNeXt attained high classification performance for music genre classification.

The salient attributes (findings, advantages, limitations, future works, practical implications) of this research are discussed below.

Findings:

•The proposed GuitarNeXt architecture, a new-generation CNN, achieves outstanding performance in music genre classification, with a test accuracy of 96.40% and an overall precision of 96.59%.

•By integrating a novel hybrid attention mechanism—combining pooling-based and convolution-based approaches—the model effectively extracts discriminative features from spectrogram images.

•The innovative downsampling and output phases, which utilize a mix of average pooling, maximum pooling, and depth concatenation, contribute to a rich feature representation while keeping the model lightweight (~4.7 million parameters).

•The staged training strategy (pretraining on ImageNet1k, fine-tuning on augmented data, and refining with rotations) further enhances the model's performance and robustness.

Advantages:

•GuitarNeXt achieves high classification accuracy and precision.

•It outperforms many traditional VGG-like CNN models.

•The model has about 4.7 million learnable parameters. This makes it efficient and suitable for resource-limited environments.

•A hybrid attention block is introduced in the model.

•Unique downsampling methods are also used.

•These innovations mark a significant step forward in CNN design for audio tasks.

•The output phase uses both global average pooling and global maximum pooling.

•This approach captures overall trends and distinctive peaks in spectrograms.

•The result is improved classification performance.

Limitations:

•The model was evaluated on a single public music genre classification dataset, which may limit its generalizability across different audio or multi-modal datasets.

•The performance under real-world conditions with noisy or highly variable audio inputs remains to be thoroughly tested.

•Further validation is required to assess the model's effectiveness on a broader range of musical genres and recording conditions.

Future directions:

•Test the GuitarNeXt architecture on larger, more diverse music genre datasets to assess its generalizability and robustness.

•Investigate combining CNNs with transformer architectures or recurrent layers to capture temporal dependencies in music.

•Optimize the model for real-time music classification applications on edge devices and mobile platforms.

•Extend the architecture to related audio tasks such as speech recognition or environmental sound classification.

Practical implications:

•Integrating GuitarNeXt into music streaming platforms can lead to more accurate genre classification, enhancing recommendation systems.

•The model can be applied to organize large music libraries, automatically tagging and categorizing tracks by genre.

•Its lightweight design makes it suitable for deployment on consumer electronics, such as smartphones and embedded systems, for on-device music classification.

•The architectural innovations can serve as a blueprint for developing next-generation deep learning models in audio signal processing, benefiting both academic research and commercial applications.

6. CONCLUSIONS

GuitarNeXt is a new CNN for music genre classification. It achieved 96.40% test accuracy and 96.59% overall precision. The model shows strong performance in classifying music. It uses about 4.7 million parameters. The model is lightweight and efficient. It works well on devices with limited resources. One innovative feature is the hybrid attention block. It combines pooling-based and convolution-based methods. This block extracts clear and useful features from spectrogram images. The downsampling phase is unique. It uses a mix of average pooling, maximum pooling, and depth concatenation. This approach solves the routing problem without extra convolution layers. The architecture has four phases. It includes a patchify-based stem, a dedicated GuitarNeXt phase, a downsampling phase, and an output phase. This design improves overall performance. The training strategy is staged. It includes pretraining on ImageNet1k, fine-tuning on augmented data, and refinement with rotations. This training strategy increases the model's robustness. These innovative aspects and high accuracy make GuitarNeXt a valuable tool. It can be used in music streaming, automated music library organization, and real-time audio processing.

REFERENCES

[1] Pourmoazemi, N., Maleki, S. (2024). A music recommender system based on compact convolutional transformers. Expert Systems with Applications, 255: 124473. https://doi.org/10.1016/j.eswa.2024.124473

[2] Ding, Y., Zhang, H., Huang, W., Zhou, X., Shi, Z. (2024). Efficient music genre recognition using ECAS-CNN: A novel channel-aware neural network architecture. Sensors, 24(21): 7021. https://doi.org/10.3390/s24217021

[3] Duan, Y., Wang, J. (2022). Design of semiautomatic digital creation system for electronic music based on recurrent neural network. Computational Intelligence and Neuroscience, 2022(1): 5457376. https://doi.org/10.1155/2022/5457376

[4] Zhang, J. (2021). Music feature extraction and classification algorithm based on deep learning. Scientific Programming, 2021(1): 1651560. https://doi.org/10.1155/2021/1651560

[5] Cai, X., Zhang, H. (2022). Music genre classification based on auditory image, spectral and acoustic features. Multimedia Systems, 28(3): 779-791.

https://doi.org/10.1007/s00530-021-00886-3
[6] Zaman, K., Sah, M., Direkoglu, C., Unoki, M. (2023). A survey of audio classification using deep learning. IEEE Access, 11: 106620-106649. https://doi.org/10.1109/ACCESS.2023.3318015
[7] Cetin, O. (2023). Accent recognition using a spectrogram image feature-based convolutional neural network. Arabian Journal for Science and Engineering, 48(2): 1973-1990. https://doi.org/10.1007/s13369-022-07086-9
[8] Moysis, L., Iliadis, L.A., Sotiroudis, S.P., Boursianis, A.D., et al. (2023). Music deep learning: deep learning methods for music signal processing—A review of the state-of-the-art. IEEE Access, 11: 17031-17052. https://doi.org/10.1109/ACCESS.2023.3244620
[9] Li, T. (2024). Optimizing the configuration of deep learning models for music genre classification. Heliyon, 10(2): e24892. https://doi.org/10.1016/j.heliyon.2024.e24892
[10] Liu, B., Feng, L., Zhao, Q., Li, G., Chen, Y. (2023). Improving the accuracy of lane detection by enhancing the long-range dependence. Electronics, 12(11): 2518. https://doi.org/10.3390/electronics12112518
[11] Wu, Q., Sun, L., Ding, N., Yang, Y. (2024). Musical tension is affected by metrical structure dynamically and hierarchically. Cognitive Neurodynamics, 18(4): 1955-1976. https://doi.org/10.1007/s11571-023-10058-w
[12] Madarapu, S., Ari, S., Mahapatra, K.K. (2024). A deep integrative approach for diabetic retinopathy classification with synergistic channel-spatial and self-attention mechanism. Expert Systems with Applications, 249: 123523. https://doi.org/10.1016/j.eswa.2024.123523
[13] Hou, R. (2024). Music content personalized recommendation system based on a convolutional neural network. Soft Computing, 28(2): 1785-1802. https://doi.org/10.1007/s00500-023-09457-2
[14] Mirza, F.K., Gürsoy, A.F., Baykaş, T., Hekimoğlu, M., Pekcan, Ö. (2024). Residual LSTM neural network for time dependent consecutive pitch string recognition from spectrograms: A study on Turkish classical music makams. Multimedia Tools and Applications, 83(14): 41243-41271. https://doi.org/10.1007/s11042-023-17105-y
[15] Ige, A.O., Sibiya, M. (2024). State-of-the-art in 1d convolutional neural networks: A survey. IEEE Access, 12: 144082-144105. https://doi.org/10.1109/ACCESS.2024.3433513
[16] Saranti, A., Pfeifer, B., Gollob, C., Stampfer, K., Holzinger, A. (2024). From 3D point-cloud data to explainable geometric deep learning: State-of-the-art and future challenges. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14(6): e1554. https://doi.org/10.1002/widm.1554
[17] Freire, P., Srivallapanondh, S., Spinnler, B., Napoli, A., Costa, N., Prilepsky, J.E., Turitsyn, S.K. (2024). Computational complexity optimization of neural network-based equalizers in digital signal processing: A comprehensive approach. Journal of Lightwave Technology, 42(12): 4177-4201. https://doi.org/10.1109/JLT.2024.3386886
[18] Chen, Y., Sun, Y. (2024). The usage of artificial intelligence technology in music education system under deep learning. IEEE Access, 12: 130546-130556. https://doi.org/10.1109/ACCESS.2024.3459791
[19] Ahmed, M., Rozario, U., Kabir, M.M., Aung, Z., Shin, J., Mridha, M.F. (2024). Musical genre classification using advanced audio analysis and deep learning techniques. IEEE Open Journal of the Computer Society, 5: 547-467. https://doi.org/10.1109/OJCS.2024.3431229
[20] Wang, L., Zhang, M., Gao, X., Shi, W. (2024). Advances and challenges in deep learning-based change detection for remote sensing images: A review through various learning paradigms. Remote Sensing, 16(5): 804. https://doi.org/10.3390/rs16050804
[21] Nguyen, N.M., Nguyen, T.T., Tran, P.N., Lim, C.P., Pham, N.T., Dang, D.N.M. (2024). Multi-modal fusion in speech emotion recognition: A comprehensive review of methods and technologies. Available at SSRN 5063214.
[22] Adamczyk, M., Oleksinski, A., Siakała, S., Swietek, J., Wierzejska, K. (2024). Automatic synthetic music generation using machine learning techniques based on user descriptions. Politechnika Śląska.
[23] Levin, D., Singer, G. (2023). Graph-based automatic feature selection for multi-class classification via mean simplified silhouette. arXiv preprint arXiv:2309.02272. https://doi.org/10.48550/arXiv.2309.02272
[24] Ba, T.C., Le, T.D.T., Van, L.T. (2025). Music genre classification using deep neural networks and data augmentation. Entertainment Computing, 53: 100929. https://doi.org/10.1016/j.entcom.2025.100929
[25] Elbir, A., Aydin, N. (2020). Music genre classification and music recommendation by using deep learning. Electronics Letters, 56(12): 627-629. https://doi.org/10.1049/el.2019.4202
[26] Wu, J., Guan, D., Yuan, W. (2025). Echo lite voice fusion network: Advancing underwater acoustic voiceprint recognition with lightweight neural architectures. Applied Intelligence, 55(2): 112. https://doi.org/10.1007/s10489-024-06035-3
[27] Oguike, O.E., Primus, M. (2025). Multimodal music genre classification of sotho-tswana musical videos. IEEE Access, 13: 28799-28808. https://doi.org/10.1109/ACCESS.2025.3536026
[28] Wang, H., SalmiJamali, S., Chen, Z., Shang, Q., Ran, L. (2022). An intelligent music genre analysis using feature extraction and classification using deep learning techniques. Computers and Electrical Engineering, 100: 107978. https://doi.org/10.1016/j.compeleceng.2022.107978
[29] Yang, J., Shi, R., Wei, D., Liu, Z., et al. (2023). Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data, 10(1): 41. https://doi.org/10.1038/s41597-022-01721-8
[30] Jin, J., Zhou, S., Li, Y., Zhu, T., Fan, C., Zhang, H., Li, P. (2025). Reinforced collaborative-competitive representation for biomedical image recognition. Interdisciplinary Sciences: Computational Life Sciences, 17(1): 215-230. https://doi.org/10.1007/s12539-024-00683-2
[31] Aytac, O., Senol, F.F., Tuncer, I., Dogan, S., Tuncer, T. (2025). An innovative approach to parasite classification in biomedical imaging using neural networks. Engineering Applications of Artificial Intelligence, 143: 110014. https://doi.org/10.1016/j.engappai.2025.110014
[32] Yılmaz, A.A. (2025). A novel deep learning-based framework with particle swarm optimisation for intrusion detection in computer networks. PloS one,

20(2): e0316253. https://doi.org/10.1371/journal.pone.0316253
[33] Chaieb, M., Azzouz, M., Refifa, M.B., Fraj, M. (2025). Deep learning-driven prediction in healthcare systems: Applying advanced CNNs for enhanced breast cancer detection. Computers in Biology and Medicine, 189: 109858. https://doi.org/10.1016/j.compbiomed.2025.109858
[34] Upadhyay, A., Chandel, N.S., Singh, K.P., Chakraborty, S.K., et al. (2025). Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture. Artificial Intelligence Review, 58(3): 92. https://doi.org/10.1007/s10462-024-11100-x
[35] Zhang, N., Ni, S., Chen, L., Wang, T., Chen, H. (2025). High-throughput and energy-efficient FPGA-based accelerator for all adder neural networks. IEEE Internet of Things Journal, 12: 20357-20376. https://doi.org/10.1109/JIOT.2025.3543213
[36] Meyers, V., Hefenbrock, M., Gnad, D., Tahoori, M. (2025). Leveraging neural trojan side-channels for output exfiltration. Cryptography, 9(1): 5. https://doi.org/10.3390/cryptography9010005
[37] GTZAN dataset - Music genre classification. https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification.
[38] Naidu, P., Rao, B.J., Revathi, K., Gayathri, M.G., Jayasri, G.S. (2025). Classification of music genres using multimodal deep learning technique. E3S Web of Conferences, 616: 02012. https://doi.org/10.1051/e3sconf/202561602012
[39] Jiang, L., Yang, L., Azimi, S. (2025). Enhanced capsule neural network with advanced triangulation topology aggregation optimizer for music genre classification. Scientific Reports, 15(1): 41. https://doi.org/10.1038/s41598-024-83577-z
[40] T.M, V., T.R, S. (2024). Music genre classification using convolution temporal pooling network. Multimedia Tools and Applications, 84: 25597-25612. https://doi.org/10.1007/s11042-024-20163-5
[41] SuriyaPrakash, J., Kiran, S. (2022). Obtain better accuracy using music genre classification systemon gtzan dataset. In 2022 IEEE North Karnataka Subsection Flagship International Conference (NKCon), Vijaypur, India, pp. 1-5. https://doi.org/10.1109/NKCon56289.2022.10126991
[42] Wijaya, N.N., Setiadi, D.R.I.M., Muslikh, A.R. (2024). Music-genre classification using Bidirectional long short-term memory and mel-frequency cepstral coefficients. Journal of Computing Theories and Applications, 1(3): 243-256. https://doi.org/10.62411/jcta.9655
[43] Srivastava, N., Ruhil, S., Kaushal, G. (2022). Music genre classification using convolutional recurrent neural networks. In 2022 IEEE 6th Conference on Information and Communication Technology (CICT), Gwalior, India, pp. 1-5. https://doi.org/10.1109/CICT56698.2022.9997961
[44] Shah, M., Pujara, N., Mangaroliya, K., Gohil, L., Vyas, T., Degadwala, S. (2022). Music genre classification using deep learning. In 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, pp. 974-978. https://doi.org/10.1109/ICCMC53470.2022.9753953
[45] Ceylan, H.C., Hardalaç, N., Kara, A.C. (2021). Automatic music genre classification and its relation with music education. World Journal of Education, 11(2): 36-45.
[46] Anuraj, K., Poorna, S.S., Renjith, S. (2024). Music Genre classification-a holistic approach employing multiple features. In 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, pp. 1-5. https://doi.org/10.1109/ICCCNT61001.2024.10725629
[47] Shen, J., Xiao, G. (2024). Music genre classification based on functional data analysis. IEEE Access, 12: 185482-185491. https://doi.org/10.1109/ACCESS.2024.3512874
[48] Murugan, B.S., Dhanasekaran, S., Kumar, P. (2024). Utilizing spectrograms and deep learning techniques for improved music genre classification. In 2024 7th International Conference on Circuit Power and Computing Technologies (ICCPCT), Kollam, India, pp. 1723-1728. https://doi.org/10.1109/AISP61711.2024.10870721
[49] Srinivas, U.M., Rafi, S., Manohar, T.V., Rao, M.V. (2024). Classification of music genre using deep learning approaches. In 2024 4th International Conference on Artificial Intelligence and Signal Processing (AISP), VIJAYAWADA, India, pp. 1-5. https://doi.org/10.1109/AISP61711.2024.10870721
[50] Chaudary, E., Aziz, S., Khan, M.U., Gretschmann, P. (2021). Music genre classification using support vector machine and empirical mode decomposition. In 2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC), Karachi, Pakistan, pp. 1-5. https://doi.org/10.1109/MAJICC53071.2021.9526251
[51] Raval, M., Dave, P., Dattani, R. (2021). Music genre classification using neural networks. International Journal of Advanced Research in Computer Science, 12(5): 12-18. https://doi.org/10.26483/ijarcs.v12i5.6771
[52] Chen, S.H., Huang, W.T., Lai, C.H., Lin, Y.L., Su, M.H. (2024). Analysis and discussion of feature extraction technology for musical genre classification. In 2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Hsinchu City, Taiwan, pp. 1-4. https://doi.org/10.1109/O-COCOSDA64382.2024.10800024
[53] Jena, K.K., Bhoi, S.K., Mohapatra, S., Bakshi, S. (2023). A hybrid deep learning approach for classification of music genres using wavelet and spectrogram analysis. Neural Computing and Applications, 35(15): 11223-11248. https://doi.org/10.1007/s00521-023-08294-6