Kanpur Philosophers
ISSN 2348-8301
International Journal of humanities, Law and Social Sciences
Published biannually by New Archaeological & Genological Society
Kanpur India

NAAGS KANPUR

Vol. X, Issue I(K) : 2023

MUSIC GENRES CLASSIFICATION USING VISION TRANSFORMERS

Aparna Panigrahy, Assistant Professor, Department of Information Technology, Nirmala Memorial Foundation College of Commerce and Science Kandivali East, Mumbai

Abstract—Music genre classification is an important task that has numerous applications, such as music recommendation systems, music search engines, and playlist generation. Traditional music genre classification techniques rely on hand-crafted features, which can be time-consuming and may not capture the full complexity of music. However, the recent introduction of vision transformers (ViT) has demonstrated remarkable performance in natural language processing (NLP) and computer vision tasks. Therefore, it is also worth exploring its potential in music genre classification. This article introduces vision transformers for music genre categorization. The spectrogram of an audio stream is represented as an image in this proposed method, and it is then fed into the ViT model for feature extraction and classification. On the widely-used GTZAN dataset, we compared our method to several state-of-the-art approaches and determined its efficacy. Our experiments show that our proposed method significantly improves accuracy, achieving an overall accuracy of 83% on the GTZAN dataset. This suggested technique shows the potential of vision transformers in music genre categorization and might lead to music recommendation system research.

Keywords—Music Genre Classification, Spectrogram, Vision Transformer (ViT)

I. Introduction

One of the most fundamental and integral aspects of people's daily Music is a universal language that has the power to evoke emotions, inspire creativity, and connect people across different cultures and generations. Music can be classified into different genres based on attributes such as melody, rhythm, instrumentation, and cultural context. Music genre classification is essential in music analysis, recommendation systems, and personalized music streaming services [1]. Music genre classification is challenging because music is a complex and dynamic signal that varies in time and frequency domains. Traditional approaches to music genre classification involve feature extraction, such as Mel-Frequency Cepstral Coefficients (MFCCs), and classification using machine learning models, such as Support Vector Machines (SVMs) or Random Forests. However, these approaches have limitations in capturing the complex relationships between the different attributes of music genres and require manual feature engineering. Recent advancements in deep learning have shown promise in improving the accuracy and efficiency of music genre classification systems. Deep learning models, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer models, can automatically learn hierarchical representations of music signals and classify them into different genres with high accuracy [2]. It is important to note that music does not have genes in the traditional biological sense. Music is a complex and dynamic signal composed of various attributes such as melody, harmony, rhythm, instrumentation, and cultural context. The task of music genre classification involves identifying the predominant characteristics of a piece of music and classifying it into a specific genre based on those attributes. However, there are several challenges and difficulties in accurately classifying music genres, including:

- Subjectivity: Music genres can be subjective and open to interpretation. For example, a particular song may be classified as "pop" or "rock," depending on the listener's interpretation.
- Variability: Music can vary in instrumentation, melody, rhythm, and cultural context. This variability can make it challenging to develop classification models that accurately classify music across various genres and styles.
- Data availability: Labeled datasets for training classification models can be limited, particularly for less popular or niche music genres. This can make it challenging to develop accurate and generalizable classification models.
- Complexity: Music is a complex signal that varies in time and frequency domains. Developing effective feature extraction methods and classification models that can capture the complex relationships between different attributes of music genres can be challenging.

Addressing these challenges and difficulties requires careful consideration of the data, the classification features, and the classification model choice. There is ongoing research in music genre classification to address these challenges and improve the accuracy and efficiency of classification models. Numerous scholars have employed diverse, intelligent methodologies for music genre classification. N. Pelchat [3] reviewed machine learning techniques in the music genre classification task. They used images of spectrograms generated from time slices of songs as the input into an NN to classify the songs into their respective musical genres. In [4], a prediction method based on the deep learning algorithm was proposed, which has the advantage of refining the music classification's correctness, precision, and effectiveness. B. Kumaraswamy [5] proposed a new music genre classification model that included two major processes: Feature extraction and classification. In the feature extraction phase, features like "non-negative matrix factorization (NMF) features, Short-Time Fourier Transform (STFT) features, and pitch features" are extracted. The extracted features are then classified via the Deep Convolutional Neural Network (DCNN) model. B Jaishankar *et al.* [6] applied an African buffalo optimization algorithm to select the best features to classify the music data for the benchmark datasets. An overall average accuracy of 82% is achieved when used with GTZAN, ISMIR, and Latin Music datasets. Y. H. Cheng *et al.* [7] applied CNN combined with Recurrent Neural Network (RNN) architecture to implement a music genre classification model. In this study, the proposed CRNN model architecture achieves an accuracy rate of 43%. M. Ashraf *et al.* [8] presented a novel hybrid model for music classification that combines the strengths of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). The proposed model was evaluated on a publicly available dataset of music signals and compared with other state-of-the-art music classification models. K. K. Jena *et al.* [9] proposed a deep learning-based hybrid model for analyzing and classifying different music genre files. The proposed hybrid model mainly uses a combination of multimodal and transfer learning-based models for classification. The results conclude that the proposed hybrid model performs better with 81% and 71% accuracy using GTZAN and Ballroom datasets, respectively, compared to other models. R. Yang *et al.* [10] demonstrated a hybrid architecture, the PRCNN, to improve the performance of music genre classification. This end toend learning architecture consists of parallel CNN and BiRNN blocks for feature extraction. The results show that all the CNNs with a parallel RNN block performed better than CNNs alone. S. K. Prabhakar [11] applied five interesting and novel approaches for music genre classification, such as the proposed Weighted Visibility Graph-based Elastic Net Sparse Classifier (WVG-ELNSC), the proposed classification using sequential machine learning analysis with Stacked Denoising Autoencoder (SDA) classifier, the proposed Riemannian Alliance based Tangent Space Mapping (RA-TSM) transfer learning techniques, classification using Transfer Support Vector Machine (TSVM) algorithm, and finally the proposed deep learning classifier with Bidirectional Long Short-Term Memory (BiLSTM) cum Attention model with Graphical Convolution Network (GCN) termed as BAG deep learning model is used here. A. Kumar in [12] evaluated the performance of stacked auto-encoder-based deep neural networks for designing a speech/music classifier on S&S and GTZAN datasets using visual features. The best classification accuracy of 93.05% and 94.73% is observed for fused features for S&S and GTZAN datasets, respectively. W. Hongdan*et al.* [13] presented a new approach, including feature extraction and classification, that considers the disparities in spectrums. The results are obtained based on the parameters of the accuracy of 97%, precision of 94%, recall of 86.5%, F-1 score of 77.8%, and an average loss of audio signal of 40% for the proposed technique. The automatic music genre classification methods are not good at dealing with various data distributions with significant intraclass differences. Based on the issue, in [14], authors proposed an effective and parameter-efficient structure the locally activated gated neural network. LGNet significantly outperforms the existing methods for music genre classification, achieving superior performance on the filtered GTZAN dataset.

These are our primary contributions, in brief:

- At first, we pre-processed, augmented, and transformed the raw music signals, then converted those into Mel-spectrogram images.
- A comparative study was conducted to find the best ViT model for the classification task.
- It can be noted that the classification accuracy was significantly improved, and the convergence loss was minimum due to the transformer methods.

The following are the other sections of the article: The paper's proposed strategy is presented in Section II, the result discussion with performance evaluation is covered in Section III, and the conclusion with some potential future directions is discussed in Section IV.

II. Methodology

The vision transformer architecture used to categorize music into many genres is described in this section. Finding the most appropriate, accurate vision transformer model is the primary goal of this research. Here are the general steps required for music genre classification using vision transformers:

- Pre-processing: The audio signals are first pre-processed by converting them into spectrograms, two-dimensional representations of the audio signals. The spectrograms are then normalized and resized to a fixed size to ensure that they can be processed efficiently by the vision transformer model.
- Model training: The vision transformer model is trained on the pre-processed spectrograms, where the model learns to extract meaningful features from the input data. The training process involves feeding the spectrograms into the vision transformer model, which applies a series of transformer blocks to generate high-level features for classification.
- Feature extraction: Once the model is trained, it can be used for feature extraction. The spectrogram is extracted and fed into the trained vision transformer model for a given audio signal to obtain high-level features.
- Genre classification: The extracted features are then used to classify the music genre using a feed-forward neural network classification algorithm.
- Evaluation: The performance of the classification algorithm is evaluated using metrics such as accuracy, precision, recall, and F1 score. The assessment is usually performed on a test set of audio signals not used during the training phase.

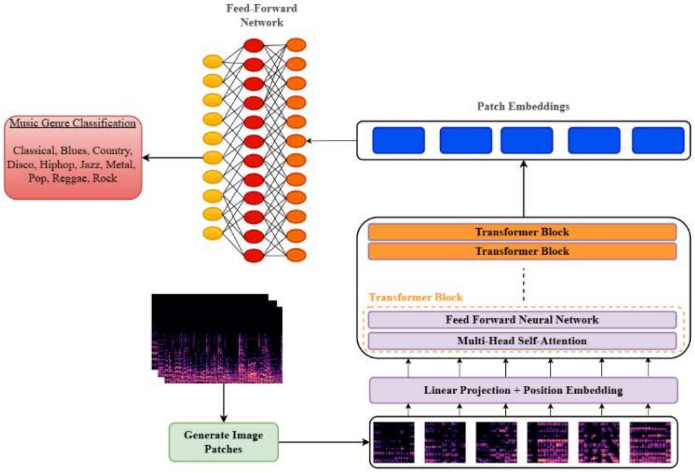
Feed-Forward Network
Music Genre Classification
Classical, Blues, Country, Disco, Hiphop, Jazz, Metal, Pop, Reggae, Rock
Patch Embeddings
Transformer Block
Transformer Block
Transformer Block
Feed Forward Neural Network
Multi-Head Self-Attention
Linear Projection + Position Embedding
Generate Image Patches

Fig. 1. General architecture of Vision Transformer (ViT)

A. *Vision Transformer Architecture*

The basic architecture of vision transformer (ViT) models consists of two main components: the transformer encoder and the classification head, shown in Fig.1. In ViT models, the input is typically a two-dimensional image divided into fixed size patches. The patches are then flattened into a sequence of vectors and passed through the transformer encoder. The transformer encoder learns to extract meaningful features from the input patches, which are then aggregated to produce a global feature representation of the image. The global feature representation is then used for classification by the classification head.

B. *Data Acquisition*

The GTZAN dataset consists of 1000 audio recordings, each lasting 30 seconds. There are 100 tracks from each of the ten genres. The covered genres include jazz, reggae, pop, rock, classical, blues, disco, country, hip-hop, metal, and blues. The .wav extension is used for the audio files. The Mel-Spectrograms created from the .wav files are then used as input for the ViT model. Out of the 1000 images, 800 (80 × 10) are selected for training, and 200 (20 × 10) are isolated for testing. In a spectrogram, the frequency components of a signal are graphically shown in relation to time. Other names for spectrograms include sonographs, voiceprints, and voicegrams. When the spectrogram is magnified to 3D, they are also called waterfalls. Spectrograms are studied (STFT) using band-pass filters or the short-time Fourier transform. This study makes use of the Mel spectrogram, another sort of spectrogram. The spectrogram has been transformed into a Mel scale. Each genre's .wav audio signals are shown in Fig.2, and the corresponding Mel Spectrogram is demonstrated in Fig.3.

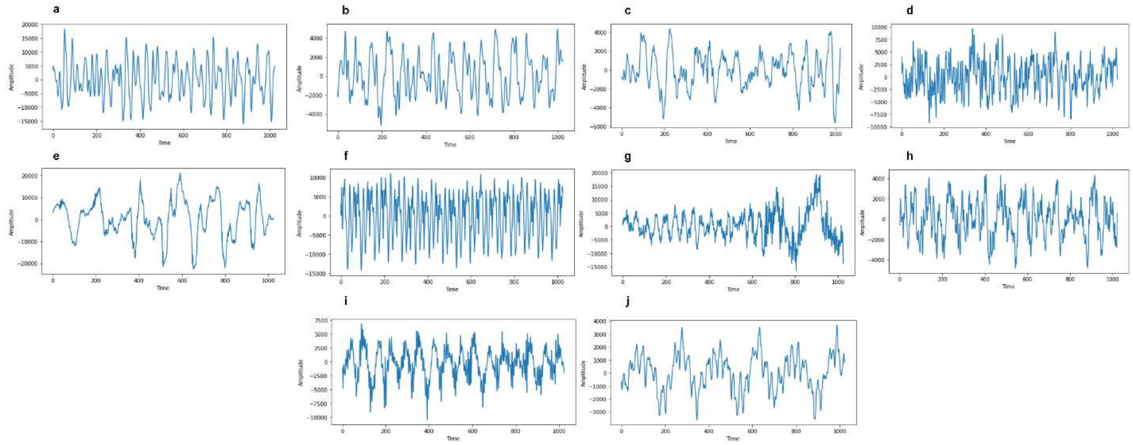

Fig. 2. Different music genre signals of GTZAN dataset. (a) Blues, (b) Classical, (c) Country, (d) Disco, (e) Hip-hop, (f) Jazz, (g) Metal, (h) Pop, (i) Reggae, and (j) Rock

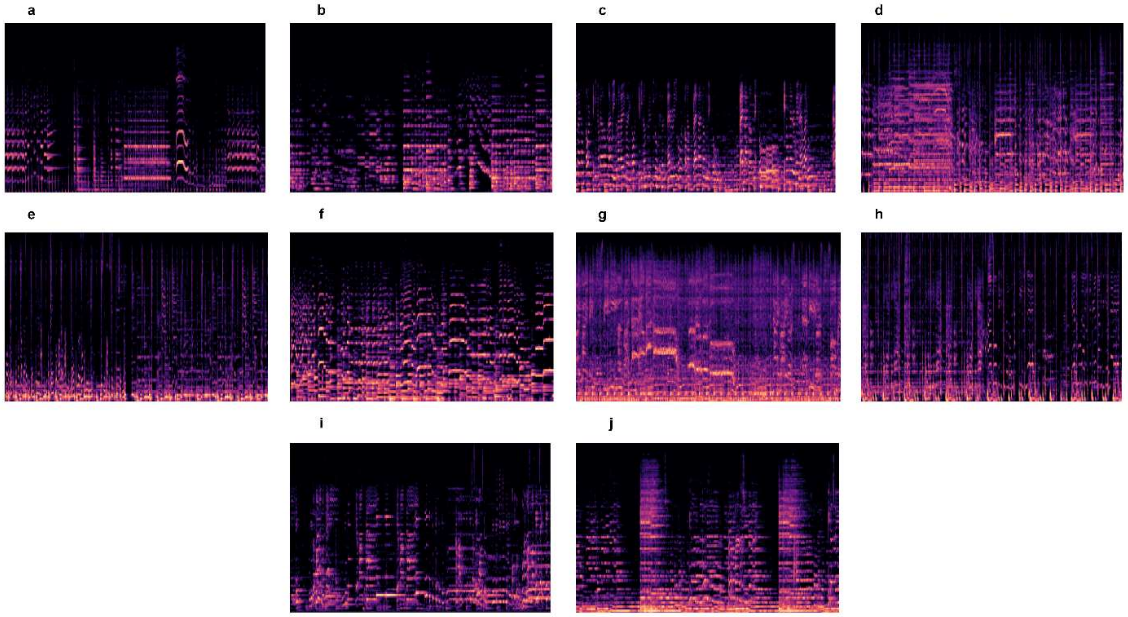

Fig. 3. Different music genre Mel-spectrogram signals of GTZAN dataset. (a) Blues, (b) Classical, (c) Country, (d) Disco, (e) Hip-hop, (f) Jazz, (g) Metal, (h) Pop, (i) Reggae, and (j) Rock

C. Data Training

The training procedure for a vision transformer (ViT) model involves the following steps:

- Data preparation: The first step is to prepare the training data. The training data should be labeled images in a format that the ViT model can read. The images should be pre-processed, normalized, and resized to a fixed size.
- Model architecture selection: The next step is to select the architecture of the ViT model. The architecture includes the number of transformer blocks, the number of attention heads, the size of the hidden layers, and other hyperparameters. The architecture can be selected based on the complexity of the task and the available computational resources.
- Initialization: The ViT model is initialized with random weights. The weights can be initialized using various methods, such as Xavier or He.
- Training loop: The training loop consists of the following steps:
  - Forward pass: The input image is passed through the ViT model to obtain high-level features. The high level features are obtained by applying a series of transformer blocks to the input image. The output of the last transformer block is the global feature representation of the image.
  - Loss computation: The global feature representation of the image is then passed through a linear layer to obtain a set of class probabilities. The class probabilities are compared to the ground truth labels using a loss function, such as cross-entropy loss. The loss measures the difference between the predicted class probabilities and the ground truth labels.
  - Backward pass: The gradients of the loss concerning the model parameters are computed using backpropagation. The gradients are then used to update the model parameters using an optimizer, such as Adam or SGD.
- Hyperparameter tuning: During the training process, various hyperparameters, such as the learning rate, batch size, and regularization, can be tuned to improve the performance of the ViT model.
- Validation: To monitor the performance of the ViT model during training, a validation set can be used to compute the accuracy or other metrics. The validation set is typically a set of labeled images that are not used during training.

- Early stopping: To prevent overfitting, early stopping can stop the training process when the performance on the validation set starts to degrade.
- Testing: Once the ViT model is trained, it can be tested on labeled images not used during training or validation. The test set is used to evaluate the performance of the ViT model on unseen data.

All the experiments were conducted on a computer with an operating system Windows 11 OS, NVIDIA RTX 4090, with 24 GB and 64 GB RAM GPU memory. The Python= 3.10, CUDA 11.7, and cuDNN 8.4.0 accelerated environments, along with the Visual Studio 2022 framework for parallel computing and deep neural network library, were utilized to carry out the training procedure.

D. Evaluation

The loss and accuracy curves are used to evaluate how well the suggested models work. The curves may be generated using the statistical indices from the confusion matrix in the forms of False Positive (FP), True Positive (TP), True Negative (TN), and False Negative (FN), as previously mentioned. Lastly, performance metrics, including specificity, recall, F1- score, Fowkes-Mallows index (FM), error rate, accuracy, Cohen's kappa coefficient (k), and precision, are evaluated based on 1- 9.

$$Accuracy = \left(\frac{TP+TN}{Numberoftotaldata}\right) \times 100\% \quad (1)$$

$$Recall= \left(\frac{TP}{TP+FN}\right) \times 100\% \quad (2)$$

$$Specificity = \left(\frac{TN}{FP+TN}\right) \times 100\% \quad (3)$$

$$Precision = \left(\frac{TP}{TP+FP}\right) \times 100\% \quad (4)$$

$$F1\text{-}score = 2 \times \left(\frac{Precision \times Recall}{Precision+Recall}\right) \quad (5)$$

$$Error\ Rate = \left(\frac{FP+FN}{Numberoftotaldata}\right) \times 100\% \quad (6)$$

$$MCC = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (7)$$

$$FM = \sqrt{\frac{TP}{TP+FP} \times \frac{TP}{TP+FN}} \quad (8)$$

$$k = \frac{2*(TP*TN-FN*FP)}{(TP+FP)*(FP+TN)+(TP+FN)*(FN+TN)} \quad (9)$$

III. Results & Discussion

This multi-class music genre classification task trains and tests the model on 800 and 200 Mel-spectrogram pictures. This paper employed Google's ViT model for the classification task and compared it with Microsoft's Swin transformer and Facebook's Convnext model. Table I shows the testing image accuracy for each of the ViT models. Table II presents the confusion matrix corresponding to the model that produced the best results. Table III displays the performance indicators, including MCC, FM, recall, precision, specificity, F1-score, accuracy, etc. It is observed that using the swin transformer V2 model, the overall accuracy of 83% was achieved. The radar map depiction of the performance indices for each musical genre is shown in Fig. 4.

Table. I. Accuracy of the Various ViT Models

ViT Models	**Accuracy (%)**
Google's ViT	75
Microsoft's Swin Transformer	75.50
Microsoft's Swin Transformer v2	83
Facebook's Convnext	77
Facebook's Convnext v2	77.50

Table. II. Confusion Matrix of the GTZAN Dataset

True Class	**Predicted Class**									
	Blues	Classical	Country	Disco	Hip-hop	Jaza	Metal	Pop	Reggae	Rock
Blues	18	0	0	1	0	1	1	0	0	1
Classical	1	18	0	0	0	0	0	0	0	1
Country	1	0	13	1	0	0	0	1	0	1
Disco	1	0	0	19	2	0	0	1	0	1
Hip-hop	0	0	0	0	17	0	0	0	0	0
Jazz	0	0	0	1	0	20	0	0	0	1
Metal	0	0	0	0	0	0	20	1	0	0
Pop	0	0	0	2	1	0	0	15	0	1
Reggae	1	0	1	0	1	0	0	0	16	0
Rock	2	0	1	1	0	0	2	2	1	10

Table. III. Performance Evaluation Parameter of GTZAN Dataset with the Swin Transformer v2

Music Genres	**Performance Parameters (%)**								
	Accuracy	Precision	Recall	F1-score	Specificity	Error Rate	MCC	FM	k
Blues	95	75	81.81	77.88	96.62	5	76.12	78.33	78.44
Classical	99	100	90	94.73	100	1	94.34	94.86	94.18
Country	97.5	86.66	81.25	83.86	98.91	2.5	82.77	83.91	82.51
Disco	94.5	76	79.16	77.54	96.59	5.5	75.13	77.56	74.41
Hip-hop	98	80.95	100	89.47	97.81	2	88.98	89.97	88.38
Jazz	98.5	95.23	90.90	93.01	99.43	1.5	92.26	93.03	92.18
Metal	98	86.95	95.23	90.90	98.32	2	89.98	90.99	89.78
Pop	95.5	75	78.94	76.91	97.23	4.5	75.03	76.94	74.43
Reggae	98	88.04	84.21	88.83	99.44	2	88.04	88.97	87.79
Rock	97.5	62.5	52	56.76	96.68	2.5	54.99	57	53.06
Overall Percentage (%)	97.15	62.63	83.35	82.98	98.1	2.85	81.76	83.15	81.51

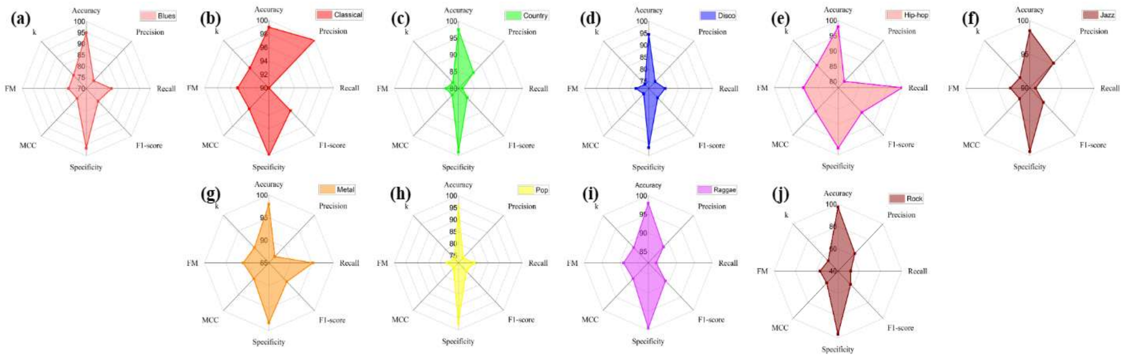

Fig. 4. Music Genre Performance Representation using. (a) Blues, (b) Classical, (c) Country, (d) Disco, (e) Hip-hop, (f) Jazz, (g) Metal, (h) Pop, (i) Reggae, and (j) Rock

IV. Conclusion & Future Directions

In conclusion, music genre classification using vision transformers is a promising area of research that has the potential to improve the accuracy and efficiency of music genre classification systems. Using vision transformers, we can transform audio signals into visual representations that deep learning models can process to classify music genres accurately. The results of the current studies in this area have shown that vision transformers can outperform traditional deep-learning models for music genre classification tasks. Using transformer encoders allows the model to capture long range dependencies in the audio signals and extract meaningful features relevant to the classification task. The classification head then maps the features to a set of class probabilities, which can be used to classify the music genres. However, there is room for improvement in music genre classification using vision transformers. One potential improvement area is using more extensive datasets and diverse music genres to train the models. This can help improve the models' generalization ability and make them more robust to variations in music styles and instrumentation. Another potential area of improvement is the development of novel pre-processing techniques for audio signals that can improve the performance of vision transformers. This can include methods for feature extraction, normalization, and data augmentation. In the future, we can expect to see further advancements in the field of music genre classification using vision transformers, with the potential for new applications in music recommendation systems, audio-based search engines, and personalized music streaming services.

References

[1]. C. Plut, P. Pasquier, "Generative music in video games: state of the art, challenges, and prospects", *Entertainment Computing*, Vol. 33, 2020.

[2]. K. Palanisamy, Kamalesh, D. Singhania, and A. Yao, "Rethinking CNN models for audio classification", *arXiv preprint* arXiv:2007.11154, 2020.

[3]. N. Pelchat, & C. M. Gelowitz, "Neural network music genre classification," *Canadian Journal of Electrical and Computer Engineering*, vol. 43, no.3, pp. 170-173, 2020.

[4]. W. Zhang, "Music Genre Classification Based on Deep Learning", *Mobile Information Systems*, 2022.

[5]. B. Kumaraswamy, and P. G. Poonacha, "Deep convolutional neural network for musical genre classification via new self adaptive sea lion optimization", *Applied Soft Computing*, 2021.

[6]. B. Jaishankar, R. Anitha, F. D. Shadrach, M. Sivarathinabala, & V. Balamurugan, "Music Genre Classification Using African Buffalo Optimization," *Computer Systems Science And Engineering*, vol. 44, no. 2, pp. 1823-1836, 2023.

[7]. Y. H. Cheng, P. C. Chang,D. M. Nguyen, & C. N. Kuo,"Automatic Music Genre Classification Based on CRNN," *Engineering Letters*, vol. 29, no.1, 2020.

[8]. M. Ashraf, F. Abid, I. U. Din, J. Rasheed, M. Yesiltepe, S. F. Yeo, & M. T. Ersoy, “A Hybrid CNN and RNN Variant Model for Music Classification,” *Applied Sciences*, vol. 13, no. 3:1476, 2023.

[9]. K. K. Jena, S. K. Bhoi, S. Mohapatra, & S. Bakshi, “A hybrid deep learning approach for classification of music genres using wavelet and spectrogram analysis,” *Neural Computing and Applications*, vol. 35, pp 1-26, 2023.

[10]. R. Yang, L. Feng, H. Wang, J. Yao, & S. Luo, “Parallel recurrent convolutional neural networks-based music genre classification method for mobile devices,”*IEEE Access*, vol. 8, pp. 19629-19637, 2020.

[11]. S. K. Prabhakar, and S.W. Lee, “Holistic Approaches to Music Genre Classification using Efficient Transfer and Deep Learning Techniques”, *Expert Systems with Applications*, 2022.

[12]. Kumar, S. S. Solanki, and M. Chandra, “Stacked auto-encoders based visual features for speech/music classification”, *Expert Systems with Applications*, 2022.

[13]. W. Hongdan, S. SalmiJamali, C. Zhengping, S. Qiaojuan, and R. Le, “An intelligent music genre analysis using feature extraction and classification using deep learning techniques”, *Computers and Electrical Engineering*, 2022.

[14]. Z. Liu, T. Bian, M. Yang, “Locally Activated Gated Neural Network for Automatic Music Genre Classification,” *Applied Sciences*, vol. 13, no.8:5010, 2023.