IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Music Genre Classification Using Deep Learning Based Pre-Trained Convolutional Neural Networks

Aparna Panigrahy

Nirmala Memorial Foundation College of Commerce and Science

Kandivali East, Mumbai, India

Abstract—The vast field of audio signal processing, and music information retrieval includes fascinating research areas such as music genre detection. The main goal of the music recommendation playlist is to choose a collection of songs that belong to the same genre. The majority of research during the last two decades concentrated on traditional classifiers employing feature-based machine learning techniques. Due to the manual extraction of features having a significant negative impact on classification accuracy, multi-class classification difficulties, and handling large amounts of data, the existing approaches suffer from a number of drawbacks. This paper shows that ImageNetPretrained deep CNN models can be used as solid baseline networks for audio classification. Here, the GTZAN dataset is used to classify the music into ten genres with the help of deep learning-based convolutional neural networks. The experiment on multi-label music genre categorization is then used to assess the effectiveness of the suggested strategy. The metrics used to determine the findings include the suggested technique's overall accuracy of 99.5%, precision of 98%, and error rate of 0.5%.

Index Terms—Music Genre Classification, Spectrogram, Transfer Learning, Convolutional Neural Network (CNN)

I. Introduction

One of the most fundamental and integral aspects of people's daily life is music. Additionally, it may significantly enhance a person's emotional, physical, spiritual health, and mental by assisting in the eradication of bad emotions and sentiments like despair, loneliness, and depression [1]. Various music genres are different from each other, which makes people have different music preferences. A music genre is a term used to describe a form of music that has been influenced by historical or cultural roots. It may also refer to specific techniques or instrument types [2]. Theoretical comprehension of the methods utilised to form genres and the mechanisms by which humans grasp the differences between multiple genres may benefit from research in music genre categorization. Another aspect is that a person's taste in music may reveal a lot about their personality and the cultural characteristics of a place [3]. Since understanding the actual form of the music needs in-depth prior knowledge, all music media platforms use text labels for music classification or retrieval. As a result, automating the process of identifying musical tags enables the development of interesting content for both users and content producers, such as playlist creation and music discovery [4].

Some of the typical difficulties encountered in this music genre classification include the fact that, occasionally, the validation findings require further refinement. Additionally, data augmentation should take appropriate factors into account. When suggesting new approaches for music genre classification, factors like data quantity, outstanding and indepth feature assessment, a better regularisation element of the model, and improvement in the complicated network structure must be properly considered [5]. In light of these factors, transfer learning frameworks, deep learning frameworks, and sparse representation frameworks appear to be more effective and adaptable than traditional machine learning approaches in the current environment.

In the classification of musical genres, some of the wellknown as well as more recent connected works are explored. For speech/music classification (SMC), researchers have experimented with a wide range of features and classifiers throughout the years, with encouraging results. However, the majority of the compositions are built on commonplace elements. Short-time energy (STE), Zero crossing rate (ZCR), spectral flux, spectral centroid, spectral entropy, spectral rolloff, and Mel frequency cepstral coefficients are some of the common temporal and spectral parameters for building SMC (MFCC). Using CNN and Support Vector Machine (SVM), the five classes of the standard GTZAN dataset's music genre categorization were provided, with an accuracy of 78% [6]. T. Gong [7] built a Deep Belief Networks (DBN)based multifeatured fusion music classification method on a public dataset from a music website, reporting an 82.23% classification accuracy. Recurrent neural networks (RNNs) with channel attention mechanisms were employed by some well-known and recent deep learning projects to categorise the musical genres for the GTZAN dataset, and they reported a

classification accuracy of 93.1% [8]. Both deep learning and metaheuristic models were used to accurately classify musical

genres, as demonstrated by Yang [9] who used a swarmbased neural network model for music genre classification and Chen [10] who used Hidden Markov Models (HMM) for music genre classification. Kumaraswamy & Poonacha [11] used a deep CNN model with a new Self-Adaptive Sea Lion
Optimization for the classification task.

These are our primary contributions, in brief:

- We demonstrated that the pre-trained neural networks could do the task of music genre classification with proper tuning of the hyper-parameters.
- At first, we did the pre-processing, augmentation, and transformation of the raw music signals, then converted those into Mel-spectrogram images.
- It can be noted that the classification accuracy was significantly improved, and the convergence loss was minimum due to the transfer learning method.

The other sections of the article are as follows: Section II presents the recommended strategy, Section III presents performance evaluation, and Section IV concludes the paper with possible future directions.

II. Methodology

This section describes the pre-trained neural network architecture, which is used for multi-genre classification of music. The main objective of this research is to discover the most appropriate pre-trained CNN model with the highest accuracy. The entire method is separated into three fundamental phases, which are explained below: data acquisition, training of the data, and evaluation.

A. Data Acquisition

A collection of 1000 audio recordings, each lasting 30 seconds, is included in the GTZAN dataset. 100 tracks from each of the ten genres are included. Classical, blues, disco, country, hip-hop, metal, jazz, reggae, pop, and rock are some of the genres that are represented. The audio files are in the form of .wav extensions. The .wav files are converted into Mel-Spectrograms, which are then given as input to the Convolutional Neural Network (CNN). Out of 1000 images, 800 (80×10) are selected for training, and 200 (20×10) are separated for testing.

The frequency contents of a signal as they relate to time are represented visually in a spectrogram. Sonographs, voiceprints, and voicegrams are some other names for spectrograms. They are also known as waterfalls when the spectrogram is expanded to 3D. Utilizing band-pass filters or the short-time Fourier transform, spectrograms are analysed

TABLE I
Workstation Specifications

Components	Description
OS	Windows 10, 64 bits
Processor	Intel(R) Core (TM) i7-8700 CPU @ 3.20 GHz
Memory	64 GB
Graphics	NVIDIA Quadro P4000 (8 GB)

(STFT). Here, another type of spectrogram is used in this work, known as the Mel spectrogram. It is a spectrogram that is converted to a Mel scale. Fig.1 shows the .wav audio signals of each genre, and Fig.2 shows the Mel Spectrogram image of the same music genre.

B. Data Training

The next step is to use Transfer Learning to transfer knowledge from one or more domains to a new domain with a different goal task. In this process, the untrained CNNs are trained with ImageNet weights. After that, the pre-trained output layers, such as the fully connected layer, softmax layer, and output layer, were replaced with layers containing the multi classes of the GTZAN dataset, and this process is called fine-tuning. Fig.3 shows the general architecture of a CNN with transfer learning. In this research we have used 18 pretrained neural networks namely Densenet 121, Densenet 169, Densenet 201, Inception v3, InsepectionResnet v2, Mobilenet, Mobilenet v2, NasnetMobile, NasnetLarge, Resnet 50, Resnet 101, Resnet 152, Resnet 50v2, Resnet 101 v2, Resnet 152 v2, Vgg 16, Vgg 19, and Xeception.

It takes a lot of computing power to train modern CNN models. As a result, all the tests were performed on a workstation, with the specifications in Table I. The training process was conducted with the help of Python= 3.8.13, CUDA 11.1, and cuDNN 8.0.5; accelerated environments were used, with Visual Studio 2019 framework for parallel computing and deep neural network library.

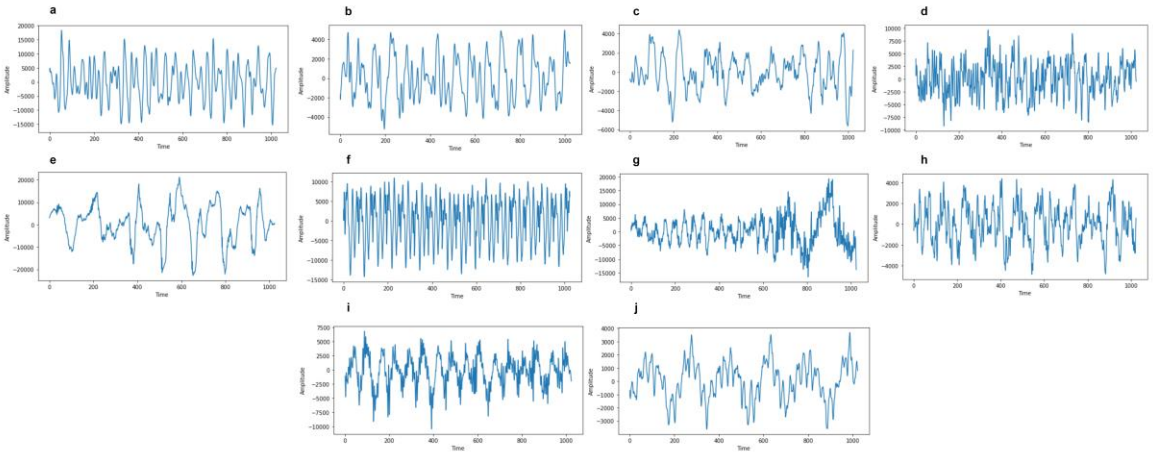

Fig. 1. Different music genre signals of GTZAN dataset. (a) Blues, (b) Classical, (c) Country, (d) Disco, (e) Hip-hop, (f) Jazz, (g) Metal, (h) Pop, (i) Reggae, and (j) Rock

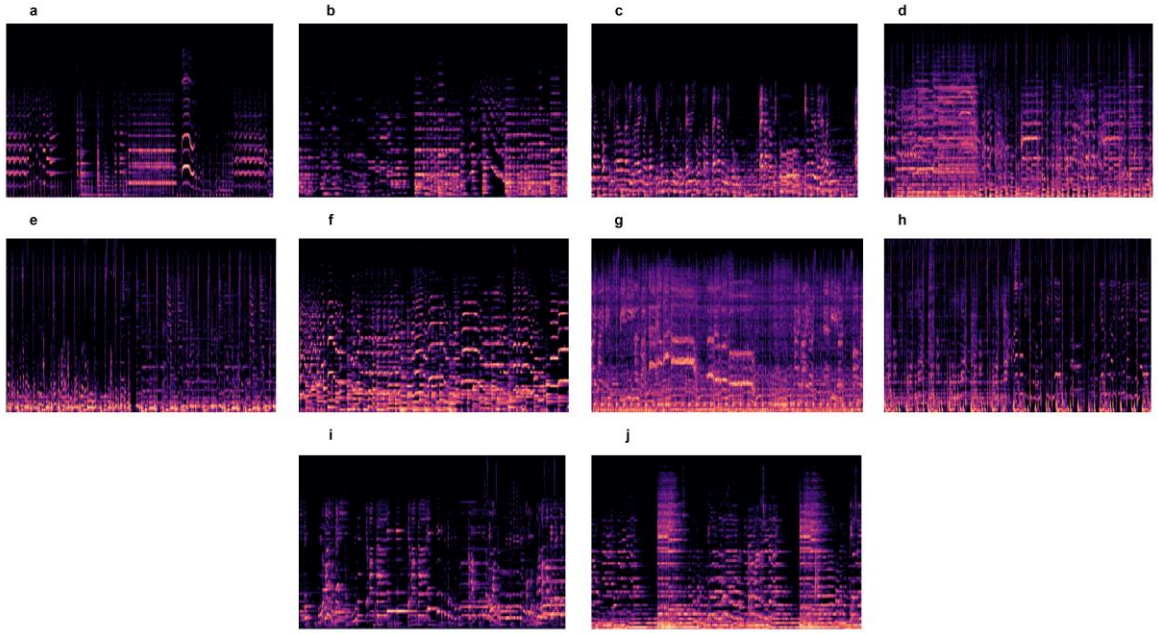

Fig. 2. Different music genre Mel-spectrogram signals of GTZAN dataset. (a) Blues, (b) Classical, (c) Country, (d) Disco, (e) Hip-hop, (f) Jazz, (g) Metal, (h) Pop, (i) Reggae, and (j) Rock

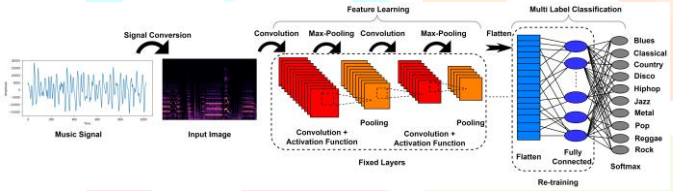

Fig. 3. General architecture of pre-trained convolutional neural networks.

C. Evaluation

The loss and accuracy curves are used to determine the suggested models' performances. The curves, as mentioned earlier, can be obtained by calculating the statistical indices from the confusion matrix in the form of False Positive (F_P), False Negative (F_N), True Positive (T_P), and True Negative (T_N). Finally, the performance indices like precision, accuracy, error rate, recall, F1-score, specificity, Fowlkes-Mallows index (FM), and Matthews correlation coefficient (MCC) are evaluated based on 1- 8.

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \times 100\% \quad (1)$$

$$ErrorRate = \frac{F_P + F_N}{T_P + T_N + F_P + F_N} \times 100\% \quad (2)$$

$$Specificity = \frac{T_N}{F_P + T_N} \times 100\% \quad (3)$$

$$Recall = \frac{T_P}{T_P + F_N} \times 100\% \quad (4)$$

$$Precision = \frac{T_P}{F_P + T_P} \times 100\% \quad (5)$$

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\% \quad (6)$$

$$MCC = \frac{(T_P \times T_N) - (F_P \times F_N)}{\sqrt{(T_P + F_P)(T_P + F_N)(T_N + F_P)(T_N + F_N)}} \times 100\% \quad (7)$$

$$FM = \sqrt{\frac{T_P}{T_P + F_P} \times \frac{T_P}{T_P + F_N}} \times 100\% \quad (8)$$

III. Results and Discussion

In the present multi-class music genre classification task, the model is trained and tested on 800 and 200 Mel-spectrogram pictures, respectively. The 18 types of pre-trained neural networks with Adam optimizer are used to find the best accurate

TABLE II
PRE-TRAINED NEURAL NETWORK PERFORMANCE

Pre-trained CNN	Validation Accuracy (%)	Validation Loss
DenseNet 121	99.22	0.05423
DenseNet 169	99.31	0.01359
DenseNet 201	99.50	0.00623
Inception V3	95.70	0.11122
InceptionResNet V2	98.05	0.11249
MobileNet	98.44	0.38318
MobileNet V2	97.27	0.09458
NASNetMobile	98.44	0.05012
NASNetlarge	97.66	0.10088
ResNet 50	92.58	0.22019
ResNet 101	87.50	0.25305
ResNet 152	90.62	0.31827
ResNet 50 V2	98.05	0.19194
ResNet 101 V2	97.66	0.08548
ResNet 152 V2	98.83	0.03326
VGG 16	99.22	0.04553
VGG 19	96.88	0.06889

Xception	98.05	0.05668

and most suited model. Table II displays the validation loss and testing image accuracy for each of the 18 pre-trained CNNs. The confusion matrix of the model with the best outcome is shown in Table III. In addition, desired outcomes characteristics can be shown in terms of the performance and optimization learning curves, which are depicted in Figs. 4 and 5. Table IV displays the different performance indicators including MCC, recall, FM, specificity, precision, F1-score, accuracy, etc. It can be observed that using DenseNet 201 with transfer learning and Adam optimizer, the overall accuracy of 99.5% has been achieved. The comparison of the current approach with various deep learning models is shown in Table V.

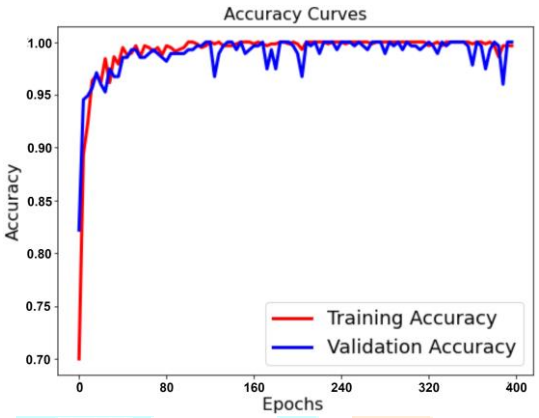

Fig. 4. DenseNet 201 Pre-trained CNN performance learning curve.

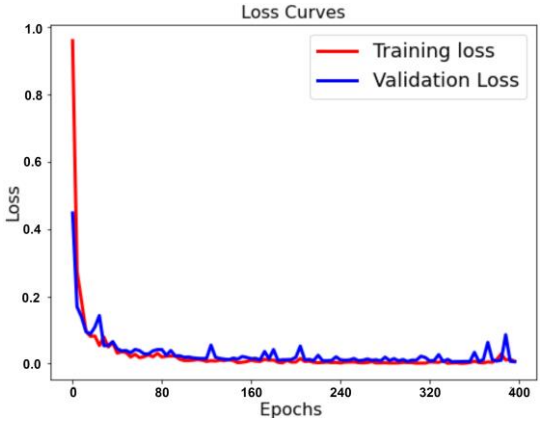

Fig. 5. DenseNet 201 Pre-trained CNN convergence curve.

TABLE V

COMPARISON OF MUSIC GENRE CLASSIFICATION WITH OTHER TECHNIQUES

Reference Number	Technique Applied	Overall Accuracy (%)
[12]	Machine Learning & Autoencoder	93.51
[13]	Autoencoder	94.73
[14]	Pre-trained CNN with BiLSTM	97
Present Method	Pre-trained CNN with Transfer Learning	99.5

TABLE III

CONFUSION MATRIX OF THE GTZAN DATASET

True Class	Predicted Class									
	Blues	Classical	Country	Disco	Hip-hop	Jazz	Metal	Pop	Reggae	Rock
Blues	20	0	0	0	0	0	0	0	0	0
Classical	0	20	0	0	0	0	0	0	0	0
Country	0	0	20	0	0	0	0	0	0	0
Disco	0	0	0	20	0	0	0	0	0	0
Hip-hop	0	0	0	0	20	0	0	0	0	0
Jazz	0	0	0	0	0	15	5	0	0	0
Metal	0	0	0	0	0	0	20	0	0	0
Pop	0	0	0	0	0	0	0	20	0	0
Reggae	0	0	0	0	0	0	0	0	20	0
Rock	0	0	0	0	0	0	0	0	0	20

TABLE IV

PERFORMANCE EVALUATION PARAMETERS OF GTZAN DATASET WITH THE BEST CNN (DENSENET 201)

Music Genres	Performance Parameters (%)							
	Accuracy	Precision	Recall	F1-Score	Specificity	Error Rate	MCC	FM
Blues	100	100	100	100	100	0	100	100
Classical	100	100	100	100	100	0	100	100
Country	100	100	100	100	100	0	100	100
Disco	100	100	100	100	100	0	100	100
Hip-hop	100	100	100	100	100	0	100	100
Jazz	97.5	100	75	85.71	100	2.5	85.26	86.60
Metal	97.5	80	100	88.88	97.22	2.5	88.19	89.49
Pop	100	100	100	100	100	0	100	100
Reggae	100	100	100	100	100	0	100	100
Rock	100	100	100	100	100	0	100	100
Overall Percentage (%)	99.5	98	97.5	97.45	99.72	0.5	97.34	97.60

IV. Conclusion and Future Directions

With the development of music streaming media technology, a playlist now contains hundreds of songs that are often categorised by musical genre. The gradual increase in the quantity of musical compositions has been attributed to the rapid development of Internet technology. As a result, the classification of music genres is a crucial component of music information retrieval. Since human labelling and annotation by music specialists is a time-consuming process, machine learning and deep learning algorithms have taken the place of the old methods used by professionals to identify the various music genres. It is commonly utilised by academics because to the simplicity of its implementation. This paper compared 18 pre-trained neural networks for the classification of music genres. The pre-trained neural networks showed a fast convergence rate and high accuracy performance. As a result, Densenet 201 achieved 99.5% overall accuracy and a loss of 0.00623 due to transfer learning. Therefore, it can be expected that the pre-trained neural networks can be implemented in IoT devices successfully. This task can be explored in the future with different optimizers and methods such as autoencoder, Vision Transformers (ViT), and Generative Adversarial Networks (GAN) models.

References

[1] C. Plut, P. Pasquier, "Generative music in video games: state of the art, challenges, and prospects", *Entertain. Comput.*, 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] S. Shin, J. Kim, Y. Yu, S. Lee, and K. Lee, "Self-Supervised Transfer Learning from Natural Images for Sound Classification", *Applied Sciences*, Vol. 11, no. 7, 3043, 2021.

[4] L. Nanni, Y. M. G. Costa, R. L. Aguiar, Jr. C. N. Silla, and S. Brahnam, "Ensemble of deep learning, visual and acoustic features for music genre classification", *Journal of New Music Research*, Vol. 47, no. 4, pp. 383397, 2018.

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

[6] G. Gwardys, and D. M. Grzywczak, "Deep image features in music information retrieval", *International Journal of Electronics and Telecommunications*, Vol. 60, no. 4, pp. 321-326, 2014.

[7] T. Gong, "Deep belief network-based multi feature fusion music classification algorithm and simulation", *Complexity*, 2021.

[8] J. Gan, "Music feature classification based on recurrent neural networks with channel attention mechanism", *Mobile Information Systems*, 2021.

[9] J. Yang, "A novel music emotion recognition model using neural network technology", *Frontiers in Psychology*, 2021.

[10] Y. Chen, "Automatic classification and analysis of music multimedia combined with hidden markov model", *Advances in Multimedia*, 2021.

[11] 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.

[12] 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.

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

[14] 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.