Neural Computing and Applications (2023) 35:11223–11248
https://doi.org/10.1007/s00521-023-08294-6

ORIGINAL ARTICLE

A hybrid deep learning approach for classification of music genres using wavelet and spectrogram analysis

Kalyan Kumar Jena[1] · Sourav Kumar Bhoi[1] · Sonalisha Mohapatra[1] · Sambit Bakshi[2]

Received: 4 March 2022 / Accepted: 6 January 2023 / Published online: 19 January 2023

Abstract
Manual classification of millions of songs of the same or different genres is a challenging task for human beings. Therefore, there should be a machine intelligent model that can classify the genres of the songs very accurately. In this paper, a deep learning-based hybrid model is proposed for the analysis and classification of different music genre files. The proposed hybrid model mainly uses a combination of multimodal and transfer learning-based models for classification. This model is analyzed using GTZAN and Ballroom datasets. The GTZAN dataset contains 1000 music files classified with 10 different kinds of music genres such as Metal, Classical, Rock, Reggae, Pop, Disco, Blues, Country, Hip-Hop and Jazz, and the duration of each music file is 30 s. The Ballroom dataset contains 698 music files classified into 8 different kinds of music genres such as Tango, ChaChaCha, Rumba, Viennese waltz, Jlive, Waltz, Quickstep and Samba, and the duration of each music file is 30 s. The performance of the model is evaluated using the Python tool. The macro-average and weighted average are taken for computing the percentage of accuracy of each model. From the results, it is found that the proposed hybrid model is able to perform better as compared to other deep learning models such as the convolution neural network model, transfer learning-based model, multimodal model, machine learning models and other existing models in terms of training accuracy, validation accuracy, training loss, validation loss, precision, recall, F1-score and support.

Keywords Music genre classification · Deep learning · Transfer learning · Multimodal

1 Introduction

Music is well known to humans since time immemorial. It is the sole credit of music that is responsible for implanting the core of emotions within humans. Music is the only medium that connects all of us in the world and plays an intensive role in a human's life. Everyone prefers to listen to different kinds of songs due to differences in their taste in music and their difference connects them in an invisible bond. Hence, it is important to classify songs on the basis of their genres so that they can be easily accessed by their users. However, the classification of sounds by analyzing millions of songs is a challenging task, and classifying songs manually is a very hectic process. Hence, it is important to classify songs as per their identifying characteristics. In the current era, music genre classification is a very profound matter of interest as there is a steep increase in various styles of music and a fast-growing pace of audience. Music genre classification is considered as a concept that helps users to identify and classify music into their respective categories or genres. Here, users can able to differentiate between two music categories based on the beats, artists, etc. From Jazz to Hip-Hop, Rock to Reggae and Metal to Pop music lovers are completely dependent on the current technology for categorizing as manually classifying mechanism is a near to impossible process.

✉ Kalyan Kumar Jena
kalyan.cse@pmec.ac.in

Sourav Kumar Bhoi
sourav.cse@pmec.ac.in

Sonalisha Mohapatra
sonalishamohapatro07@gmail.com

Sambit Bakshi
bakshisambit@ieee.org

[1] Department of Computer Science and Engineering, Parala Maharaja Engineering College (Government), Berhampur 761003, India

[2] Department of Computer Science and Engineering, National Institute of Technology (NIT), Rourkela 769008, India

In the music genre classification process [1–20], it is required to categorize the music files based on a given dataset. Manual classification of millions of songs of the same or different genres is a challenging task for human beings. Different methodologies can be used for such classification. Many such works are conducted by researchers with different models and methods. Panagakis et al. [21] proposed JSLRR (joint sparse low-rank representation) for music file genre classification. The special cases of JSLRR called as JSR (joint sparse representation) and LRR (low-rank representation) are also proposed by the authors. These three are compared with SRC (sparse representation-based classifier), LRC (linear regression classifier), SVM (support vector machine), NN (neural network) and type machine learning models. These models are trained and tested with GTZAN, ISMIR, Homburg, 1517Artists and Unique song datasets. From results, it is found that mostly JSLRR, JSR and LRR perform better by showing high accuracy than other models. Lykartsis et al. [22] proposed a method for presenting beat histogram features for musical genre classification using different novelty functions such as Baseline, Rhythmic, Combined, *P*.-MaxP and *P*.-Avg. The datasets used for this classification are GTZAN, Ballroom, ISMIR04, Unique, and Homburg. From the results, it is found that the combined function performs better than other novelty functions in showing higher accuracy. From the above study, it is found that the average accuracy is around 70% and this needs to be increased.

DL can also be considered as a better option to classify music genres. In today's era, DL has amazed one of the most effective tools for handling larger grades of data which utilizes complex algorithms. DL mechanism can take the help of neural network for categorizing a class of entities with their identifying features. It allows machines to recognize, extract and incept accurate properties of different music files. Different deep learning models such as CNN model, transfer learning-based model and the multimodal model can be used for this classification which will assist to classify audio files into various genres and putting them for training purposes. After training, the performance of the trained model can be analyzed for its performance. Along with it, the hybridized model can also be used to carry out this classification mechanism to test the classification accuracy.

The main contributions of this work are stated as follows.

- This work proposed a DL-based approach for the analysis and classification of music genre files. The DL-based approach is focused on the models, such as the CNN model, transfer learning-based model, multimodal training model, and proposed hybrid model to carry out such classification.
- The DL-based models are analyzed using the Ballroom dataset [23] and GTZAN dataset [24]. The GTZAN dataset consists of 1000 music files classified into 10 different kinds of music genres such as Metal, Classical, Rock, Reggae, Pop, Disco, Blues, Country, Hip-Hop and Jazz. The Ballroom dataset consists of 698 music files classified with 8 different kinds of music genres such as tango, ChaChaCha, Rumba, Viennese waltz, Jlive, Waltz, Quickstep and Samba.
- The performance of these models is evaluated using Python, and the performance parameters taken are training accuracy, validation accuracy, training loss, validation loss, precision, recall, F1-score and support. The macro-average and weighted average are taken for computing the percentage of accuracy of each model. The statistical test is also performed to check the performance of the proposed hybrid model.
- From the results, it is observed that the proposed hybrid model performs better with 81% and 71% of accuracy using GTZAN and Ballroom datasets, respectively, as compared to other deep learning models, machine learning models (SVM and NN), and other existing models. The computational time is also analyzed for the proposed model.

The rest of the paper is organized as follows. Section 2 describes the related works, Sect. 3 describes the methodology, and Sect. 4 describes the results and discussion. Finally, the conclusion is described in Sect. 5.

2 Related works

Different works have been carried out related to the classification of music genres. Some works are described as follows. Oramas et al. [1] anticipated that the learning of multimodal feature space enhances the overall model performance of the pure or unadulterated audio representation which is specifically pertinent while the additional model is accessible for training purposes. A methodology is also proposed during scaling down the dimensionality of the target level, which results in major enhancements in multilevel classification based on the diversity and accuracy of anticipated genres. This insinuates a deeper level of classification. Ultimately, there is a qualitative analysis of the result which shows the behavior of various models in the classification process. Feng et al. [2] indicated that class 2 and 3 classification emerges to be on the criteria of neural network. Classification accuracy is improved by creating more such datasets from the original limited music file, and its performance is also enhanced than the neural network.

Bahuleyan et al. [3] stated that the features that the traditional machine learning classifier and CNN commit a major portion toward the multiclass classification. Elbir et al. [4] focused on a music recommendation engine and music genre classification system which emphasizes on extracting emblematic characteristics which are procured by a deep neural network model is suggested. Acoustic features are extracted from this network which is further used for music genre classification.

Nanni et al. [5] put forward a group for automatic classification of music genre which in turn is a fusion of acoustic and visual (both non-handcrafted and handcrafted) characteristics taken out from audio files, and these characteristics are then evaluated, composed and finally combined to ensemble and procure a comparatively good range of classification accuracy than state-of-the-art methodology basis of music database of Latin, ISMIR 2004 and GTZAN collection of the genre. Kim et al. [6] stated that a model can employ the multimodel deep learning architecture in music genre classification; for spectrogram image of music and sequential data found CNN and recurrent neural network (RNN), they were used, respectively, with this a drop out which is representative regularizer is applied to avoid over fitting. The multimodal deep learning model stated here for found to be much more effective than unimodal deep learning for the classification of music genres. Oramas et al. [7] stated an approach for the classification of multilevel genres situated on the combinational feature embedding to gain understanding of the state-of-the-art deep learning method. A considerable amount of difference between the models is shown by the experiment which introduces a new baseline for multilabel genre classification and combines the yields for improvising results. Vishnupriya et al. [8] proposed that for training and classification proposed CNN is used for extraction of the feature which is the most important task for the analysis of audio. MFCC (MEN frequency Cepstran coefficient) is used for the sound sample's feature vector. This classifies music into many categories of genres that the expected feature vector accuracy was approximately 76% and greatly enhances the automatic classification procedure of music.

Lau et al. [9] stated that the classification of music genre with help of a DL convolutional model is opposing 5 traditional off-the-shell classifier, spectrogram and contain-based features included in feature selection. The classification was performed on the GTZAN dataset, and the accuracy of test data was 66%. Jeong et al. [10] aimed at the conventional spectral feature which learns substructure and how to reassemble it in the Cepstran modulation spectrum domain that is implemented effectively and successfully in various speech and musical application for temporary extraction of characteristical features. The result of the experiment with the help of GTZAN dataset resulted by stating the temporal feature acquired from the suggested methodology is efficient enough for procuring accuracy for classification compared to learned spectral features. Senac et al. [11] proposed a set of 8 music featured CNN model chosen along with 3 major musical dimensions such as timbre, tonality and dynamics; the CNN is trained in such a way that in the time and sequence domain the filter dimension is interpretable. The result is only 8 music features which are most efficient than 513 bins of frequency in a spectrogram and the late score fusion in between systems on the basis of two of the feature type which one 91% accuracy based on a dataset named GTZAN. Yu et al. [12] stated a newer model which incorporates an attention-based procedure on the basis of bidirectional RNN, and it deals with serial attention and parallelized attention. Parallelized attention as compared to serial attention gets better results and is more flexible. The CNN-based parallelized attention model takes SPST spectrogram as an input. Aguiar et al. [13] focused on the architecture of CNN which is majorly used for pattern recognition literature. When tasking of classification is imposed using CNN, over-fitting is the recurrent problem that happens due to insufficient training samples or because of the high dimensionality of the space. The ability to generalization is increased by exploring data augmentation techniques. The experiment held in LMD and obtained accuracy over crossed the state of the art which considers CNN enacted methodology.

Yang et al. [14] proposed a hybrid architecture named as parallel recurrent convolutional neural network (PRCNN) which refers to an end-wise network for training that combines the classification of data based on a series of time and feature extraction in a single step itself. The PRCNN and RNN blocks mainly focus on drawing out the spatial feature and temporal frame order. The output of two blocks is listed in the single depiction of data-related time series. After that into the softmax function for classification, the syncretic victories fed. It is guaranteed by the parallel structure of the network that the features drawn out are sturdy to constitute data-related time series; also the result of experiment proposed that the architecture outperforms already stated methodologies applied on the same dataset; the data oriented upon music is an instance to perform the contrasting experiment for supervised that one more PRNN block that improvises time series classification overall results as compared with utilising CNNs alone. Zhang et al. [15] stated majorly two ways for purpose of enhancing music genre classification with CNN: 1. combining loss as well as average for giving much statistical information to upper grade neural network; 2. uses shortcut connection for skipping multiple layers, basically inspired by residual learning method. The short-time Fourier transform of the audio generated signal is treated like input for CNN, and also CNN's output is again sent into another deep neural network for classification purposes. Two different network topologies are compared at the preliminary experimental stage, and results on the GTZAN dataset are shown. It is found that the abovementioned method especially the second one effectively enhances classification accuracy.

Liu et al. [16] stated that the influence of learned interaction with feature interaction for various subdivisional of ultimate categorisation results in some multifeature models. A method involving interaction related to middle-level learning feature situated on DL stated that the result of experiment shown that the design method persistently improvises the accuracy of music genre classification with 93,65 based on GTZAN dataset which in turn is superior to almost all current methods. Rajanna et al. [17] proposed a two-layer neural network and some manifold techniques for learning the classification of music genres combined with each other. The task of classification based on the public dataset accuracy rate of the deep neural network includes model sets for learning with the inclusion of SVM and '1 SVM, logistic regression and '1-regression. Combining manually prepared audio features for a genre classification task-based public dataset, the result of experiment shows that as compared to the classic learning model, the neural network represents a rich feature space. Shi et al. [18] stated a framework for neural networks on chroma feature-based classification of music genre. The chromatic feature can be represented with the time domain and frequency domain of music considering the harmonical existence; also, it is independent of some completely irrelevant features to genre classification such as timbre, volume and absolute speech and it is comparatively robust and unreactive to the noise of background and primary information such as monophonic and polyphonic distribution of music which is represented. The chroma feature-based music audio in the association with deep learning network is estimated here that this feature is fed into the VGG16 network for purpose of training and enhancing the later three layers. The classifier is then trained on the GTZAN dataset, and the result publishes that the obtained framework has a higher level of classification accuracy and better result performance. Elbir et al. [19] focused on the acquisition feature of music which is extracted with the help of digital signal processing technique and then music genre classification, and the recommendation of music genre is done by machine learning method with that deep learning method which is part of the convolutional neural network and is utilized for genres classification and recommendation of music also comparison of performances to obtain the result is studied. In this study, the GTZAN dataset is taken and the SVM algorithm is taken for the highest success. Tsaptsinos et al. [20] focused on a recurrent neural network model which helps to classify a huge dataset of intact song lyrics the lyrics represent a hierarchical layer structure in which the words constitute to form a line, the line constitutes to form a segment and the segment constitutes to form a complete song. Hierarchical attention network (HAN) is used to extract the layers and acquire the importance of word lines and segment a test model over a 117 genre dataset and reduce the dataset which is performed. Results show that HAN outperforms both the non-neural and simpler neural model and also classifies a greater range of genres than the previous research. Conclusively, HAN provides greater insights ranging from a computational perspective through the lyrical structure into language features that differentiate various musical genres. Many such works can also be referred from [25–31].

From the above analysis, it is observed that many work works have been carried out related to the classification of music genres. However, accurate classification is a challenging issue in the current scenarios. So, there is a need for the development of improved methods to carry out such classification in a better way.

3 Methodology

The proposed methodology is focused on the models such as the CNN model, transfer learning-based model, multimodal training model and a hybrid model for the classification of music genre files. These models are described as follows.

3.1 CNN model

The CNN model for the classification of music genre files is mentioned in Fig. 1. At first, the music files are labeled and converted to deep learning compatible files. For making it compactable, two varieties of visual representations such as wavelet generation and spectrogram analysis are taken into account. The wavelet and spectrogram representation is mentioned in Figs. 2 and 3, respectively. Both of these representations are generated by the librosa library. Librosa library generally refers to a python package which is used to analyze music and audio files. It assists its user's basic building blocks which is prior important to create music information retrieval systems. In this work, spectrogram analysis is focused. It is the visual

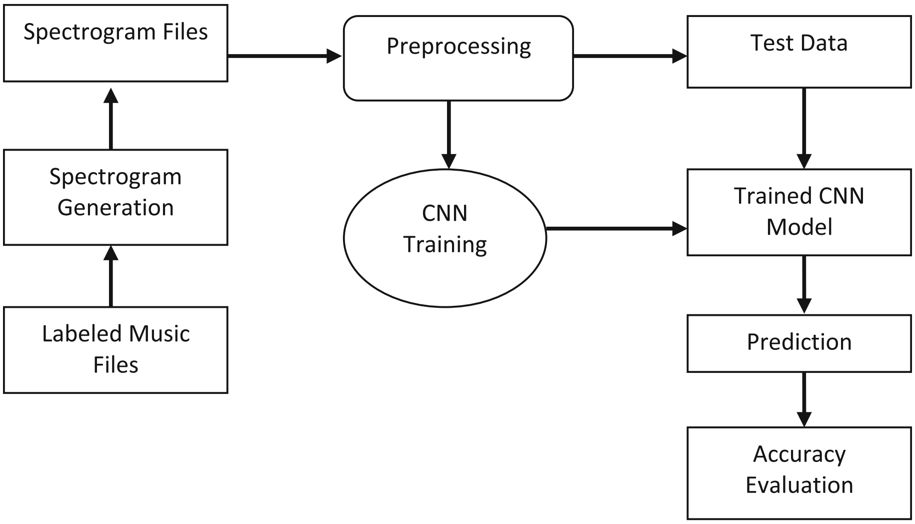


Fig. 1 CNN model for the classification of music genre files

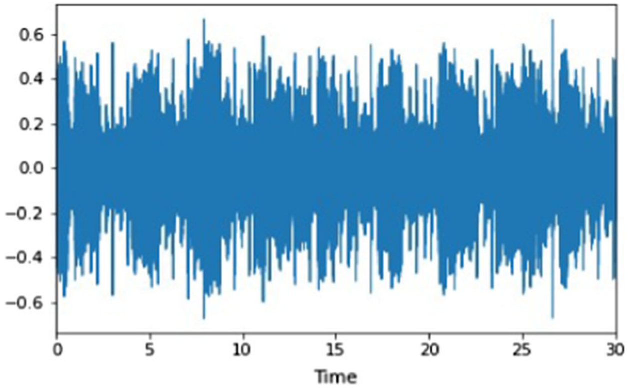


Fig. 2 Wavelet representation

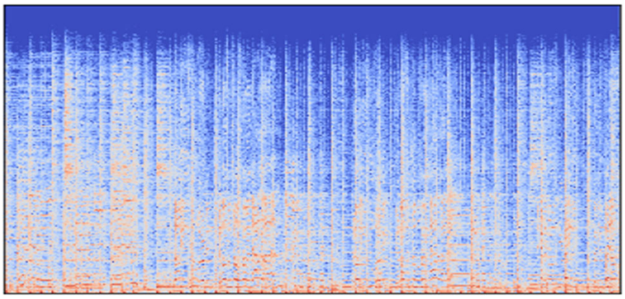


Fig. 3 Spectrogram representation

representation of frequency of waves which varies with time. However, wavelet refers to the wave-like oscillation of the amplitude of the sound. This process is used to analyze the properties of nonstationary signals like audio. After the spectrographs are created, they are kept in a file named spectrograph files. These are further saved for future testing and training purposes. Here, the librosa library is used to transform each audio file into a spectrogram. It is a package of python used to analyze musical files. It assists its user's basic building blocks which is prior important to create music information retrieval systems. After the spectrographs and waves are created, they are kept in files named wavelet files and spectrograph files. These are saved for future testing and training purposes. Preprocessing is one of the important steps as it emphasizes the data extraction process from a scratch level. Data preprocessing refers to organizing the raw and unprocessed data to make it compatible for DL further processing. After preprocessing, the test data and train data are obtained for further processing. The train data are in turn taken for training of the CNN model. Here, after preprocessing stage the train datasets are sent to train the CNN model. The dataset except the train data is considered as test data. It is used for further model training. The training dataset and test dataset together form the trained CNN model which in turn is taken for prediction and accuracy analysis. The CNN model took spectrogram data for testing and training purposes. After the generation of the trained model, the training and validation losses and model performance are measured on the basis of loss and accuracy that are checked and the precision, recall, F1-score and accuracy of the proposed model are obtained to calculate the percentage of accuracy.

The CNN model is able to feed images as input and deals with various biases in addition to different weights to different differentiable objects in a particular image. The preprocessing step in this model is less hectic as compared to other classification models. It aims to impersonate the connectivity pattern of neurons present in the brain of humans. Basically, individual neurons respond or react to stimuli strictly in the visual field within the region of restriction. Each of the collective receptive fields overlaps over each other to conquer the whole visual area. This model successfully acquires spatial and temporal dependencies by applying relevant filters in an image. This model does a better fit justice to the datasets of the image as the number of parameters involved is reduced in association with reusable weight distribution. It needs to get trained for better learning of the fact that how sophisticated an image can be. The main purpose of this model is to considerably label the images into such an accessible form which is less hectic to process, making sure that the unique critical features important for getting a good grade prediction do not get lost. It is prioritized because there is a need of designing an architecture which is more flexible for learning features as well as scalable to a big range of datasets. The CNN model normally works using the layers such as convolutional layer (CL), pooling layer (PL), RELU (rectified linear units) correction layer (RCL) and fully connected layer (FCL).

3.1.1 Convolutional layer

This layer is the basic and foundational block of ConvoNets. It is the first layer of a CNN. The convolution operation is applied to the input, and the result is passed on to the four coming layers through convolutional layers. It is the convolution itself that transforms all the pixels which are in its receptive fields into one single value and the vector is the resultant output of a convolutional layer. The functionality that the convolved layer after feeding input and passing the result to the forthcoming layer is analogous that how the visual cortex of the brain neuron responds to a particular stimulus. The data of each convolutional neuron are processed in the receptive fields only. It aims to identify the existence of a collection of characteristics for the image which is taken as input. The main principle of this filter is to brag a window which represents the featured image, and then, the product of convolution is calculated between a feature and singular portions of scanned images. This feature is again foreseen as a filter. Hence, this layer intakes various inputs as pictures, and then, the calculation of convolution operation of each filter is carried out. Then the features wanted exactly correspond with the filters. Then each pair of an image and filters is generated forming a feature map which represents that the higher value of features is feature map and is the correlation location in the picture similar that the particular characteristics.

3.1.2 Pooling layer

It acts as a sandwich layer placed between two convolutional layers and is responsible for performing pooling operations after receiving several feature maps. Basically, pooling operations include functions such as size reduction and preserving the essential characteristics. The image is sliced into regular cells with each cell containing the maximum value and avoiding the small square cells. Here, the output contains same feature map numbers in the input and it is in smaller sizes. The main purpose of this layer is to minimize some parameters amount and existing calculations inside a network, hence improvising the efficiency and avoiding over-learning. One of the biggest advantages of this pooling layer is that the values are maximum and then recognized least in terms of accuracy in the featured map which is procured thereafter in the process of pooling with that of received input.

3.1.3 RELU correction layer

RELU is a nonlinear activation function which can perform on multilayer neural networks. This layer can perform element-wise operations. It has an output which is a rectified feature map. This layer is used to perform a threshold operation on each element of input, where any value less than zero can be set to zero. This operation is represented in Eq. 1.

$$\begin{aligned} & f(t) = t, \\ & t > \ = 0 \\ & \ 0, t < \ 0 \end{aligned} \tag{1}$$

3.1.4 Fully connected layer

Every time a new output vector is generated as soon as an input vector is received. Thus, the input values are received after the application of linear combination and the feeding activation function to the values of input. This layer is responsible for the classification of the image treated as an input for a proposed network. The output is a return value of a vector consisting of the size L. Here, L is supposed to be amount of class instances in a classification problem for an image. This layer ascertains the relation between the citation of characteristics of a picture and one particular class. Here table treated as an input is the resultant of the preceding layer, and it is related to a particular feature map for a particular feature. The higher value numbers of features are indicated by the location which may be more or less accurate. The particular position of a characteristic at a fixed point in a given picture is a feature of a class; then, a particular value number in the assigned table is the already proposed significant weight.

3.1.5 Parameters used in CNN model

The CNN is differed by the layer pattern which is parameterized and stacked. The CL and PL consist of hyperparameters. Hyperparameters are the parameter values of whom values need to be defined. It is only on the hyperparameters on which the output size of feature maps of convolution is dependent. Each feature map or image is represented in Eq. 2.

$$FM = Y \times X \times N \tag{2}$$

where Y represents the width in pixels, X represents the height in pixels, N represents the number of channels and FM represents the feature map.

Generally, four numbers of hyper parameters constitute the CL. These are L (filter numbers), S (size of filters), M (the step) and O (zero–padding). For every input whose size of the image is $Y \times X \times N$, the corresponding PL outputs a matrix whose dimensions are $Yc \times Xc \times Nc$ and represented in Eq. (3).

$$\begin{aligned} Yc &= 1 + \frac{Y - S + 2O}{M} \\ Xc &= 1 + \frac{X - S + 2O}{M} \\ Nc &= L \end{aligned} \tag{3}$$

Two types of hyperparameters constitute the PL. These are Z (size of the cells) and M (step). For every input whose size of image is $Y \times X \times N$, the corresponding pooling layer outputs a matrix whose dimensions are $Yp \times Xp \times Np$ and represented in Eq. (4).

$$\begin{aligned} Yp &= 1 + \frac{W - S}{M} \\ Xp &= 1 + \frac{X - S}{M} \\ Np &= N \end{aligned} \tag{4}$$

Algorithm 1 describes the processing and classification of music genre files using the CNN model. The music genre files are processed by considering both GTZAN and Ballroom datasets. Here, the training and testing ratio is considered as 80% and 20%.

Algorithm 1. Algorithm for CNN Model

Input: Music dataset

Output- Predicting Genre of songs

```
1.Begin
2.For each audio in the dataset
3.Convert audio to spectrogram
4. For each class in the dataset
5. Train, Test data = dataset split ( 80 : 20 )
6. Training neural network with Training data
7. Prediction(Test)
8.End
```

3.2 Transfer learning-based model

The transfer learning-based model for the classification of music genre files is mentioned in Fig. 4. At first, the music files are labeled and converted to deep learning compatible files. For making it compactable, two varieties of visual representations such as wavelet generation and spectrogram analysis are taken into account. Both of these representations are generated by the librosa library. In this model, spectrogram analysis is focused. It is the visual representation of frequency of waves which varies with time. After the spectrographs are created, they are kept in a file named spectrograph files. These are further saved for future testing and training purposes. Here, the librosa library is used to transform each audio file into

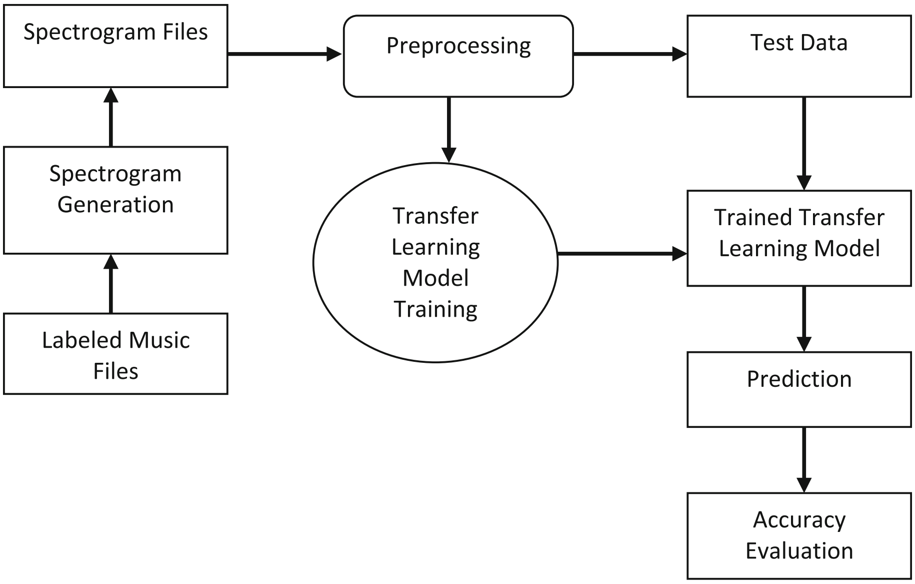


Fig. 4 Transfer learning based model for music genre classification

spectrogram. It is a package of Python used to analyze musical files. It assists its user's basic building blocks which is prior important to create music information retrieval systems. After the spectrographs and waves are created, they are kept in files named wavelet files and spectrograph files. These are saved for future testing and training purposes. Preprocessing is one of the important steps as it emphasizes the data extraction process from the scratch level. After preprocessing, the test data and train data are obtained for further processing. The train data are in turn taken for training of the model. Here, after preprocessing stage, the train datasets are then sent to model training and the process is enhanced. This model took spectrogram data for both training and testing purposes. Here a pretrained mobilet V2 model is also taken for training purposes. The dataset except the train data is considered as test data. It is used for further model training. The training dataset and test dataset together form the trained model which in turn is taken for prediction and accuracy analysis. After the generation of the trained model, the training and validation losses and model performance are measured on the basis of loss and accuracy is checked and the precision, recall, F1-score and accuracy of the proposed model are obtained to calculate the percentage of accuracy.

When humans learn new things, they always learn from scratch. The knowledge which is gained in past is also transferred and applied to newer tasks. For carrying out a particular task, one need to train models isolatedly with specific datasets for the completion of training. Also, when the tasks between the two models are correlated, the machine was incapable of transferring knowledge and insights from other model to another which in turn makes the learning procedure fully hectic. There are existing tasks in which if there is the transference of previous knowledge, then problems could be solved easily. Hence, this concept is basically used for transfer learning-based model which helps in training new models from preexisting models and is required to satisfy various tasks. This model is considered as one of the DL models. Here, a model is initially built for a specific purpose and can again be reused as an initial point of another task. It is referred to as using the parameters again and again, which were trained once on a source task to achieve a particular target task which only intends transference of knowledge between various domains. This concept emerged when there was a deficit of trained data in a target task while dealing with neural network the instances of trainable parameters of a target task model are significantly reduced by transference of pretrained weights which enhances learning from a smaller range of datasets [26]. The transfer learning is based on two main fields mainly domains and tasks. *A* domain *E* includes a space of features space and a marginal probability distribution $P(A)$, where $A = \{a1,\ldots, an\}$ € X. *A* specific domain *D* where *D* is a task which is usually composed of two main components: an objective predictive function f: $A \rightarrow b$ and a label space, and it is represented in Eq. 5.

$$D = \{A, P(A)\} \quad (5)$$

This f function is utilized for the prediction of a corresponding label $f(a)$ of a newer instance of a. Then the task is denoted by $S = \{G, f(a)\}$, which gains insights from a set of training data—consisting of pairs {ai, bi}, where a € A and bi € B. When there is a domain source Es and learning task Hs, a domain target Et and learning task Ht, where Es $\neq$ Et, or Hs $\neq$ Ht. Recently, the use of DL has steeply increased as DL models are utilized for solving many real-life problems. The problems which involve DL acquire a huge quantity of data and take a lot of time and energy for getting trained properly. Also, deep neural networks consist of millions of weights which are interlinked with an enormous number of layers of neurons together. These weights are specifically applied on inputs and are adjustable during the training phase, and the upcoming resultant is fed forward to generate output which is a very time and energy-consuming procedure. Hence, the scope for transfer learning is increased for the optimization of deep learning techniques. Transfer learning enables DL models which are pretrained on a previous problem previously used but on a newer and interrelated problem statement too hence preserving time and computational energy. Transfer learning is used in DL procedures under the circumstance such as insufficient training data. Mobile net V2 model is a residual based architecture where an input residual block and output of the residual block are in the form of thin bottleneck layers and are contrasting to old. It uses lightweight and depth-wise convolutions for filtering out characteristics from the expansion layer which is at the intermediate position. In addition to that, nonlinearity present at the narrow layers is erased for maintaining the power of representation.

Algorithm 2 describes the processing and classification of music genre files using transfer learning-based model. The music genre files are processed by considering both GTZAN and Ballroom datasets. Here, the training and testing ratio is considered as 80% and 20%.

Algorithm 2: Algorithm for Transfer Learning Based Model

Input: Music dataset

Output- Predicting Genre of songs

```
1.Begin
2.For each audio in the dataset
3.Covert audio to spectrogram
4.For each class in the dataset
5.Train, Test data = dataset split( 80 : 20 )
6.Base_Model = Mobilenetv2
7.Creating Model using Base_Model and Target Model
8.Training neural network with Training data
9.Prediction(Test)
10.End
```

3.3 Multimodal model

The multimodal model for the classification of music genre files is mentioned in Fig. 5. At first, the music files are labeled and converted to deep learning compatible files. For making it compactable, two varieties of visual representations such as wavelet generation and spectrogram analysis can be taken into account. Both of these representations are generated by the librosa library. In this model, both wavelet and spectrogram analysis is focused. After the spectrographs are created, they are kept in a file named spectrograph files. These are now further saved for future testing and training purposes. These are further saved for future testing and training purposes. Here, the librosa library is used to transform each audio file into spectrogram. It is a package of python used to analyze musical files. It assists its user's basic building blocks which is prior important to create music information

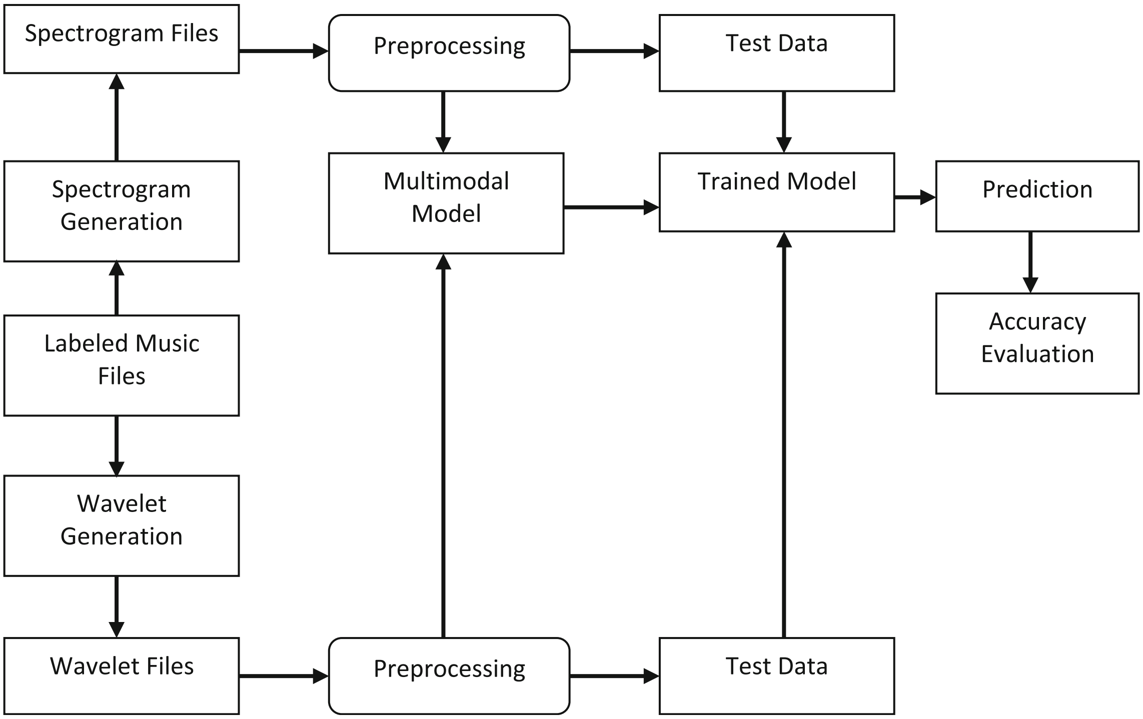


Fig. 5 Multimodal model for music genre classification

retrieval systems. After the spectrographs and waves are created, they are kept in files named wavelet files and spectrograph files. These are saved for future testing and training purposes. Preprocessing is one of the important steps as it emphasizes the data extraction process from a scratch level. After preprocessing, the test data and train data are obtained for further processing. The train data are in turn taken for training of the model. Here, after preprocessing stage the train datasets are then sent to model training and the process is enhanced. This model took spectrogram data for both training and testing purposes. Here, a pretrained mobilet V2 model is also taken for training purposes. The dataset except the train data is considered as test data. It is used for further model training. The training dataset and test dataset together form the trained model which in turn is taken for prediction and accuracy analysis. After the generation of the trained model, the training and validation losses and model performance are measured on the basis of loss and accuracy which are checked and the precision, recall, F1score and accuracy of the proposed model are obtained to calculate the percentage of accuracy. Here, the spectrogram and wavelets models are trained separately, and in later layers of convolutions, both are integrated into one training model.

The experience of humans in viewing this world is a multimodal experience as we explore objects around us, hear various things, feel the textures, smell the odors and taste various things. Basically, modality refers to the process through which something occurs or is experienced and a research problem is always categorized as multimodal when it is associated with multiple such modalities. Various kinds of modalities do have various statistical properties by category. Multimodal nets are a category of subnetworks which is accountable to convert given inputs to a joint representation form for allowance of training on training data which further includes images, sound waves, text, etc., of various sizes and dimensions. These are specifically designed for minimal computations and make sure that the majority of calculations are performed within the domain-agnostic body of the model. Multimodal nets consist of a few small modality networks, an encoder, *I*/*O* mixer and an autoregressive decoder. To get good performance across different problems, three computational blocks have to be considered. These are convolutions which make sure that the model detects local patterns and generalizes space, attention layers which focus on some specific elements to enhance the performance of the model and sparsely gated mixture of experts which provides the model capacity without excessive computation cost. This model is focused on bc(c) and bd(d) which is mentioned in Eq. (6).

$$\begin{aligned} bc(c) &= Zc2\tan yh(Zc_1c) \\ bd(d) &= Zd2\tan y(Zd_1d) \end{aligned} \tag{6}$$

where *bc*/*bd* is the formal representation of relationship embedded in a shared space. Here, c and d are the visual representations of song vectors procured from different data modalities. Zxm represents the weight matrix of *y* modality (that means *c* or *d*) of the *m*th layer. tan*y* represents the hyperbolic tangent function of each element which is combined with the nonlinear component of a network. When each song is iterated and learns two modality embeddings, then the loss function is represented using Eq. (7) and here cos(.,.) represents cosine similarity between two vectors.

$$K+ = 1 - \cos(bc(c),\, bd(d)) \tag{7}$$

The loss function for samples which are negative is mentioned in Eq. (8).

$$Kc- = \max(0, \cos(bc(uc), bd(d) - q) \tag{8}$$

For 1 modality, analogy for another part of modality is represented in Eq. (9) and here q represents the margin between scales zero and one indicating the importance of sample which is negative summarizing.

$$Kd- = \max(0, \cos(bc(c), bd(ud)) - q) \tag{9}$$

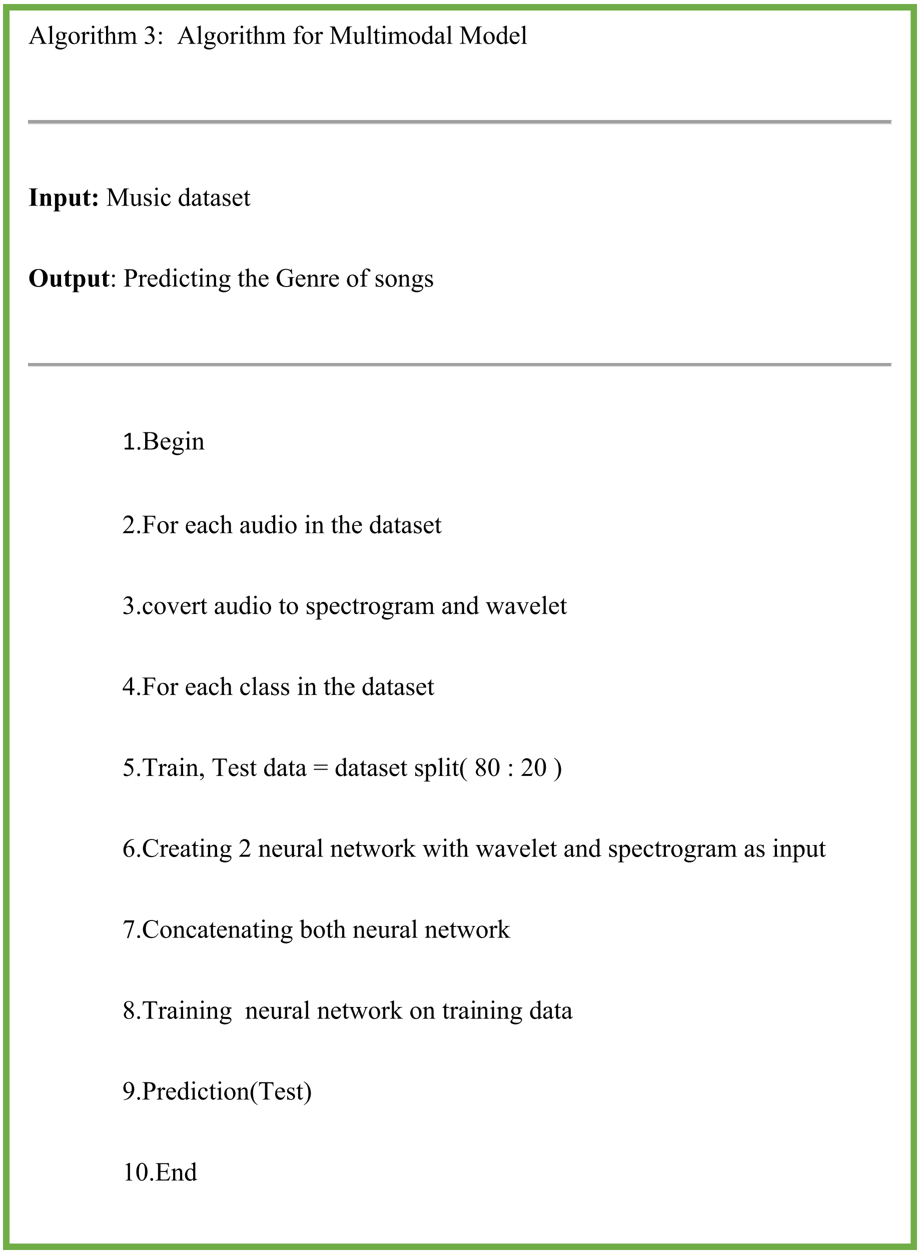
where *uc* and *ud* are the two random negative samples. The total minimal loss of a multimodal network is represented using Eq. (10) and here *bc* and *bd* each consist of two hundred dimensions.

$$K = K + +Kc - +Kd- \tag{10}$$

In multimodal fusion, different data types are combined into one. It is observed that the deep networks are enough to find optimum data text very rapidly. But due to some audio complexity, the training process is slowed down leading to under exploration of many more modalities. So, it is most important for learning each of the models separately which guarantees variable input data which is represented as a whole in each of the featured vectors. Henceforth, for any modality network, it is necessary to procure data internally to represent every training item classification of the genre. The activation of later concealed layers in every network turns the characteristics of its irrespective modality. In multilabel classification, multiple targeted labels are given to every instance which is classifiable. Mathematically, on a set of m labels $S = \{s1{,}s2\ldots.sm\}$ and a set of d items $Q = \{q1{,}q2\ldots.qd\}$ a function f is modeled which link a set of b labels to each item in *Q* where *b*€ [1,*m*] which is item variant. DL problems are best comfortable for this kind of problem as the infrastructure of deep learning allows one or more outputs in an ultimate layer. The basic infrastructure of huge multilabel classification utilizing DL usually ends up including a layer of logistic regression in the association of sigmoid activations when calculated along with cross entropy as a loss function. It is also previously assumed that the classes are statistically independent.

The most basic DL task from multimodal training is feature extraction from multimodal data fusion. Deep learning techniques these days include developing and training deep neural networks like classification and are trained on big sets of data. Intermediate features, also known as embeddings of the model which are extracted and used as input data for representation. The embeddings are more powerful as they give rich informative data as compared to rule-based or one-hot encoded vector which represents categorical data in traditional machine learning algorithms. Unimodal embeddings are used in industry, searching, clustering and categorizing data. Multimodal embeddings are utilized for downstream classification, search and clustering tasks by providing richer and better representations of data. Dense unimodal embeddings also generate multimodal embeddings producing similar results to that of multimodal embeddings. Multimodal embeddings have multiple modes of input data which informs a downstream task specifically when there are incomplete models. There are many other applications which include the visually impaired aiding, and creating precise and searchable descriptions of visual media available on the web. Visual questioning, answering and reasoning are the applications which are still under implementation. Each of the applications mentioned above faces issues for learning but it is prior important for learning to create multimodal embeddings and develop architectures. As rapidly deep learning continues to penetrate in technologies in this modern era, it is progressively important for the models to get soundly able to process multiple and very frequently incomplete sources of information too.

Algorithm 3 describes the processing and classification of music genre files using the multimodal model. The music genre files are processed by considering both GTZAN and Ballroom datasets. Here, the training and testing ratio is considered as 80% and 20%.



Algorithm 3: Algorithm for Multimodal Model

Input: Music dataset

Output: Predicting the Genre of songs

```
1.Begin
2.For each audio in the dataset
3.covert audio to spectrogram and wavelet
4.For each class in the dataset
5.Train, Test data = dataset split( 80 : 20 )
6.Creating 2 neural network with wavelet and spectrogram as input
7.Concatenating both neural network
8.Training  neural network on training data
9.Prediction(Test)
10.End
```

3.4 Proposed hybrid model

All the existing models are not able to perform better in all the cases. So, there is a need for a hybridization mechanism which can able to improve the performance of the classification work. The hybrid model is a combination of two or more models to carry out better classification mechanism. In this work, the multimodal model and transfer learning-based model (pretrained model) are hybridized to carry out the classification mechanisms. The proposed hybrid model for the classification of music genre files is mentioned in Fig. 6. At first, the music files are labeled and converted to deep learning compatible files. For making it compactable, two varieties of visual representations such as wavelet generation and spectrogram analysis can be taken into account. Both of these representations are generated by the

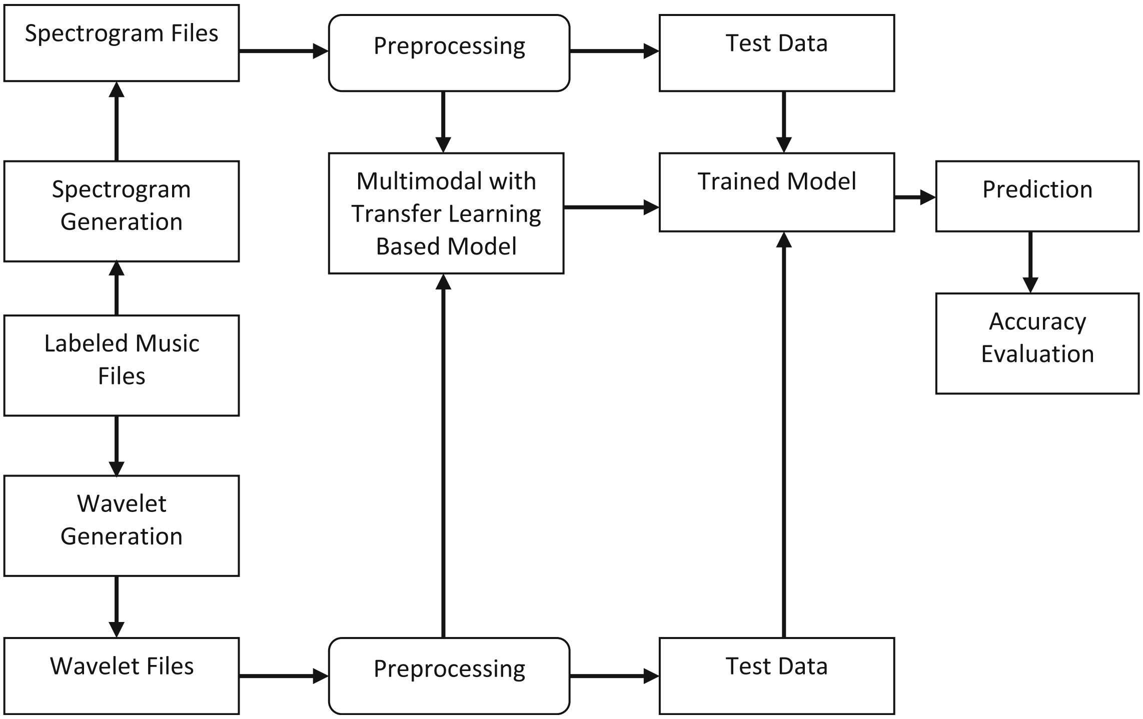
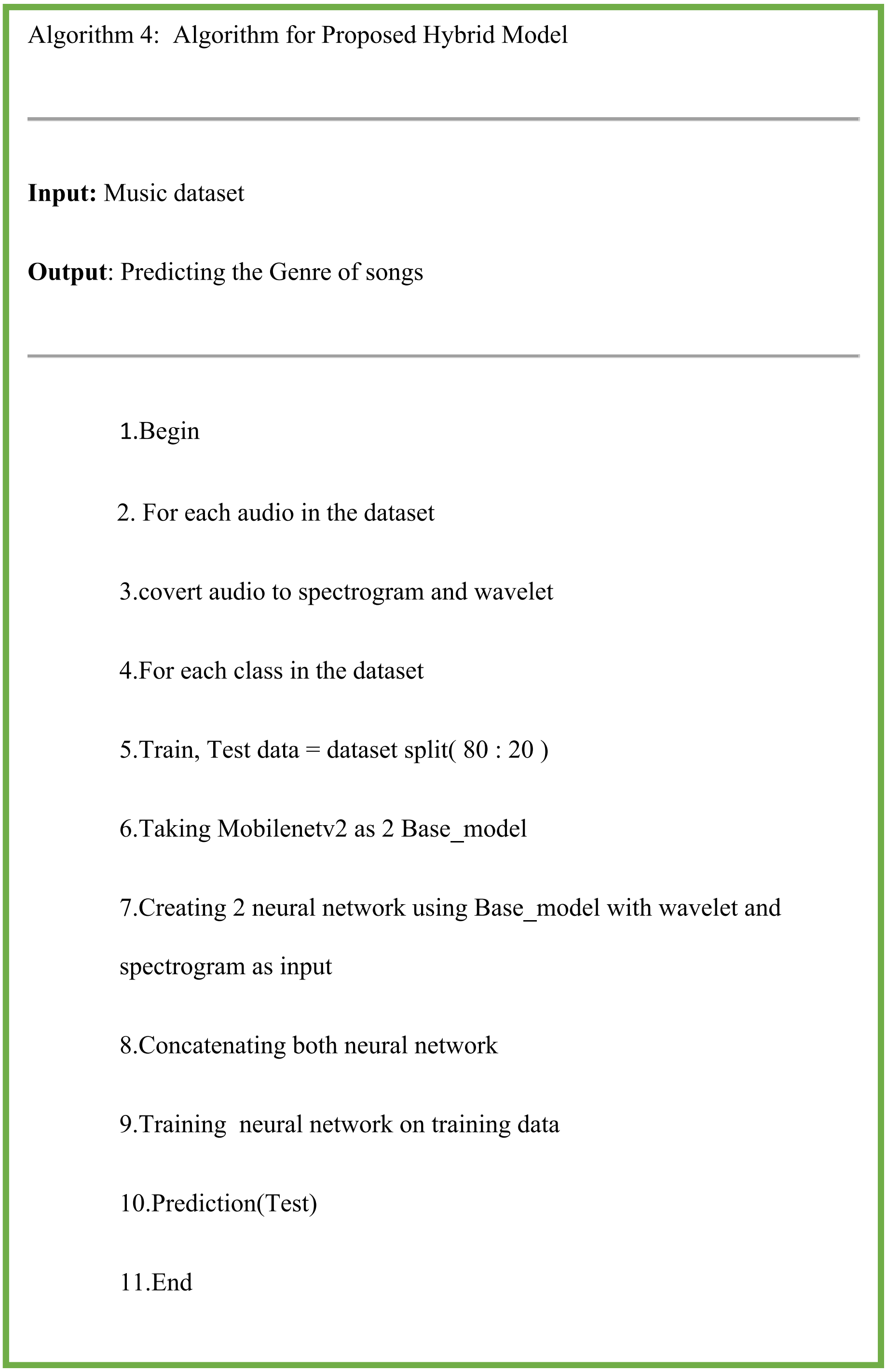


Fig. 6 Proposed hybrid model for music genre classification

librosa library. In this model, both wavelet and spectrogram analysis is focused. After the spectrographs are created, they are kept in a file named spectrograph files. These are now further saved for future testing and training purposes. These are further saved for future testing and training purposes. Here, the librosa library is used to transform each audio file into spectrogram. It is a package of python used to analyze musical files. It assists its user's basic building blocks which is prior important to create music information retrieval systems. After the spectrographs and waves are created, they are kept in files named wavelet files and spectrograph files. These are saved for future testing and training purposes. Preprocessing is one of the important steps as it emphasizes the data extraction process from a scratch level. Data preprocessing refers to organizing the raw and unprocessed data to make it compatible for DL further processing. After preprocessing, the test data and train data are obtained for further processing. The train data are in turn taken for training of the model. Here, after preprocessing stage the train datasets are then sent to hybrid model training and the process is enhanced. The hybrid model took two models for both training and testing purposes. Here the spectrogram and wavelets are trained separately, and in later layers of convolutions, both are integrated into one training model. After the generation of the trained model, the training and validation losses and model performance are measured on the basis of loss and accuracy which are checked and the precision, recall, F1-score and accuracy of the proposed model are obtained to calculate the percentage of accuracy. Here, the spectrogram and wavelets models are trained separately, and in later layers of convolutions, both are integrated into one training model.

Algorithm 4 describes the processing and classification of music genre files using the proposed hybrid model. The music genre files are processed by considering both GTZAN and Ballroom datasets. Here, the training and testing ratio is considered as 80% and 20%.



Algorithm 4: Algorithm for Proposed Hybrid Model

Input: Music dataset

Output: Predicting the Genre of songs

```
1.Begin
2. For each audio in the dataset
3.covert audio to spectrogram and wavelet
4.For each class in the dataset
5.Train, Test data = dataset split( 80 : 20 )
6.Taking Mobilenetv2 as 2 Base_model
7.Creating 2 neural network using Base_model with wavelet and
spectrogram as input
8.Concatenating both neural network
9.Training  neural network on training data
10.Prediction(Test)
11.End
```

4 Results and discussion

The proposed work has been implemented using Python programming language and its facilitating tools and libraries like matplotlib.pyplot, seaborn, keras, Sequential, Dense, Conv2D, MaxPool2D, flatten, Dropout, Image data Generator, adam, classification_report, confusion_matrix, tensorflow, cv2, os and numpy. It is because of compatibility and simplicity of Python programming language with modularity and functionality; it produces a various number of tool kits, libraries, inbuilt classes and various functions for data preprocessing, visualization and implementation.

The proposed work is carried out by focusing on the following points.

1. Labeled music files are taken from the dataset. Then they are first converted to a deep learning compactable model. For making it compactable for deep learning, we utilize two various types of visual representations that means wavelet generation and spectrogram analysis through librosa library.
2. After the spectrographs and waves are created, they are kept in files named wavelet files and spectrograph files. These are now further saved for future testing and training purposes.
3. In preprocessing, the test data and train data are hence obtained. The train data are in turn taken for training of the model.
4. Here, after preprocessing stage, the trained datasets are then sent to model training and the process is enhanced.
5. The dataset except the train data is test data. It is used for further model training.
6. The training dataset and test dataset together form our trained model which in turn is taken for prediction and accuracy analysis. The training and testing are performed using a k-fold cross-validation approach, where k represents the number of folds, and it is considered as 5 to maintain the 80:20 training and testing ratio.
7. After the trained model is created, the training losses and validation losses and model performance with accuracy and loss are checked and precision, recall, F1-score and accuracy of the proposed model are obtained to calculate the accuracy.

In this work, two datasets such as Ballroom [23] and GTZAN [24] are taken for processing. The GTZAN dataset consists of 1000 music files with a uniform distribution of 10 types of genres each. These ten types of genres are represented by a total of 100 tracks of audio files. All the tracks are Mono 22050 Hz 16-bit audio files in the format of a wave. The total size of this dataset is 1.2 GB. Each music file is 30 s long. The genres include Blues, Classical, Country, Disco, Hip-Hop, Jazz, Raggae, Rock, Metal and Pop. The Ballroom dataset contains 698 music files classified into 8 different kinds of music genres such as Tango, ChaChaCha, Rumba, Viennese waltz, Jlive, Waltz, Quickstep and Samba, and the duration of each music file is 30 s. The total size of this dataset is 1.8 GB. The system used while implementing this model should have the mentioned characteristics. The processor of the system should be Intel multicore processor-Dual Socket Intel Broadwell processor with 14 cores each. It should have vast storage of 96 GB RAM to accommodate large datasets. It should have 16 TB of secondary storage with RAID5 configured.

In this work, the CNN model, transfer learning-based model, multimodal and proposed hybrid model are implemented. Then the performance of each model is analyzed by using the performance parameters training accuracy (TA), validation accuracy (VA), training loss (TL) and validation loss (VL). Then the accuracy (percentage) is determined with the help of precision, recall F1-score and support, and on the basis of these results, the music is classified into their respective genres. The TA, VA, TL and VL are defined as follows. TA: It is the accuracy acquired if the model is applied to the training dataset. VA: It is the measure of how well the model is able to classify with the validation dataset. TL: It is the measure of how well our model is fitting the training dataset. VL: It is the measure of how well our model is fitting the new dataset.

For accuracy analysis, the following parameters are focused.

True positive (TP) When the result is actually positive and categorized as positive, then it is TP.

True negative (TN) When the result is actually negative and categorized as negative, then it is TN.

False positive (FP) When the result is actually negative and categorized as positive then it is FP.

False negative (FN) When the result is positive and categorized as negative, then it is FN.

Precision It is an ability of a model to predict correctly the positives out of all correct predictions. Mathematically, it is represented using Eq. (11).

$$\text{Precision} = \frac{TP}{(TP + FP)} \tag{11}$$

Recall It is the model's capability to correctly assume the positive output is actually positive.

F1-score It is an alternative to accuracy metrics which provides the same weightage to precision as well as recall while performance measurement with the help of accuracy. Mathematically, it is represented using Eq. (12) and here P refers to precision and R refers to recall.

$$\text{F1 - Score} = \frac{2 * P * R}{P + R} \tag{12}$$

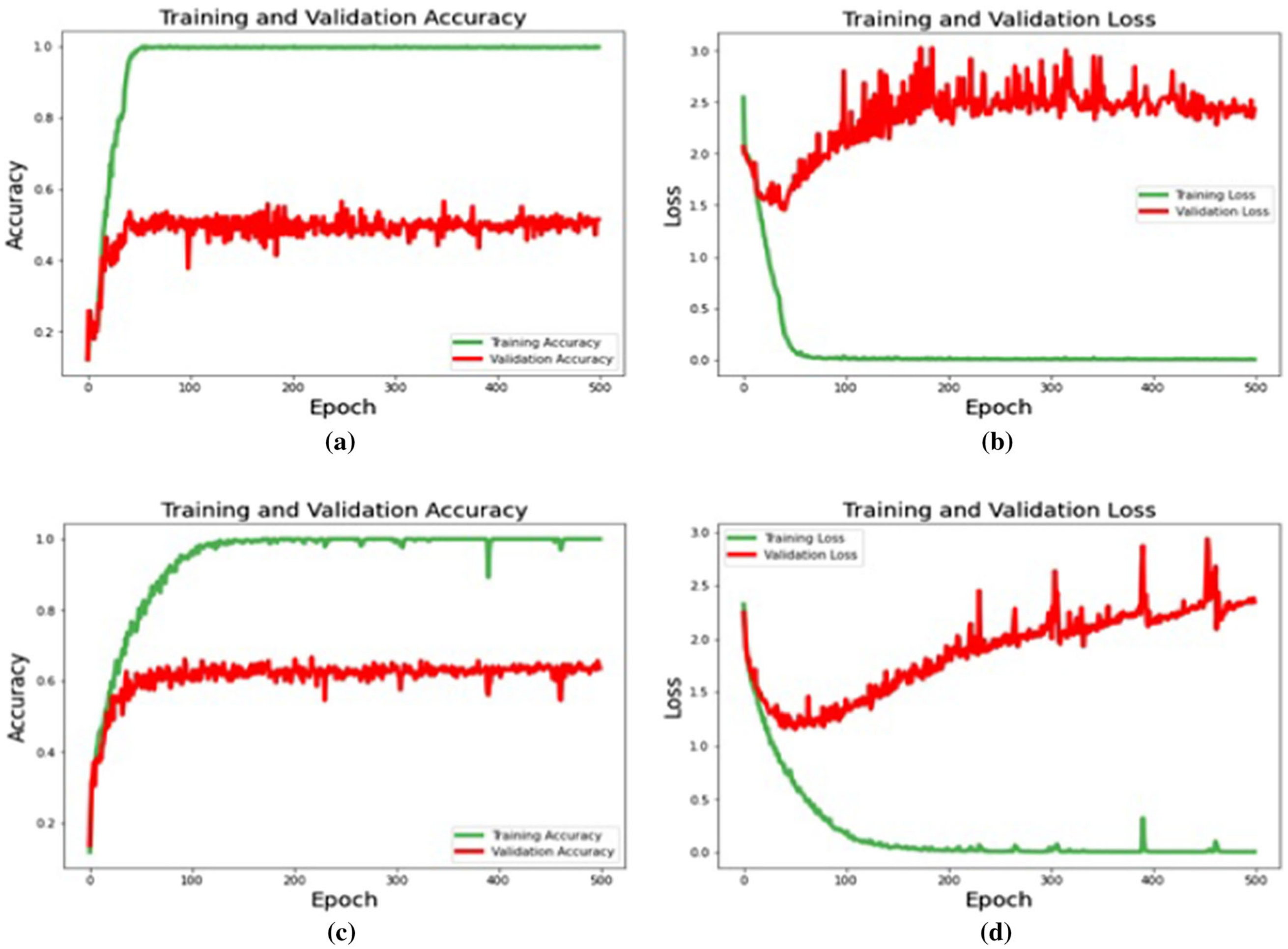
Support It is basically the total number of files present for calculating accuracy.

Accuracy It is the number of correctly categorized data instances and overall total instances. Mathematically, it is represented using Eq. (13).

$$\text{Accuracy} = \frac{TN + TP}{(TP + FN + TN + FP)} \tag{13}$$

In this work, the confusion matrix (CM) is constructed for each model. It is an $N \times N$ matrix which is utilized for performance evaluation in a classification model where N is the number of target classes. It basically correlates the actual target values with assumed values by the model, thus giving clarity on how accurate is the performance of a classification model and what types of errors it is generating. The (TA, VA) and (TL, VL) representation of the CNN model using Ballroom and GTZAN datasets is mentioned in Fig. 7. The (TA, VA) and (TL, VL) representation of the transfer learning-based model using Ballroom and GTZAN datasets is mentioned in Fig. 8. The (TA, VA) and (TL, VL) representation of multimodal model using Ballroom and GTZAN datasets is mentioned in Fig. 9. The (TA, VA) and (TL, VL) representation of the proposed hybrid model using Ballroom and GTZAN datasets is mentioned in Fig. 10.

In this work, TA refers to the accuracy obtained when the model is applied to the training data. VA refers to the accuracy obtained when the model is applied to the testing or validating data. TL is used to calculate the loss of the model on the training data, and VL is used to calculate the loss of the model on validation or testing data. Figure 7 represents the results obtained using the CNN model in terms of TA, VA, TL and VL. Here, Fig. 7a and c is focused on TA and VA representation by applying Ballroom and GTZAN datasets, respectively. Figure 7b and d is focused on TL and VL representation by applying Ballroom and GTZAN datasets, respectively. Figure 8 represents the results obtained using the transfer learning-based model in terms of TA, VA, TL and VL. Here, Fig. 8a and c is focused on TA and VA representation by applying Ballroom and GTZAN datasets, respectively. Figure 8b and d is focused on TL and VL representation by applying Ballroom and GTZAN datasets, respectively. Figure 9 represents the results obtained using the multimodal model in terms of TA, VA, TL and VL. Here, Fig. 9a and c is focused on TA and VA representation by applying Ballroom and GTZAN datasets, respectively. Figure 9b and d is focused on TL and VL representation by applying Ballroom and GTZAN datasets, respectively. Figure 10 represents the results obtained using the proposed hybrid model in terms of TA, VA, TL and VL. Here, Fig. 10a and c is focused on TA and VA representation by applying



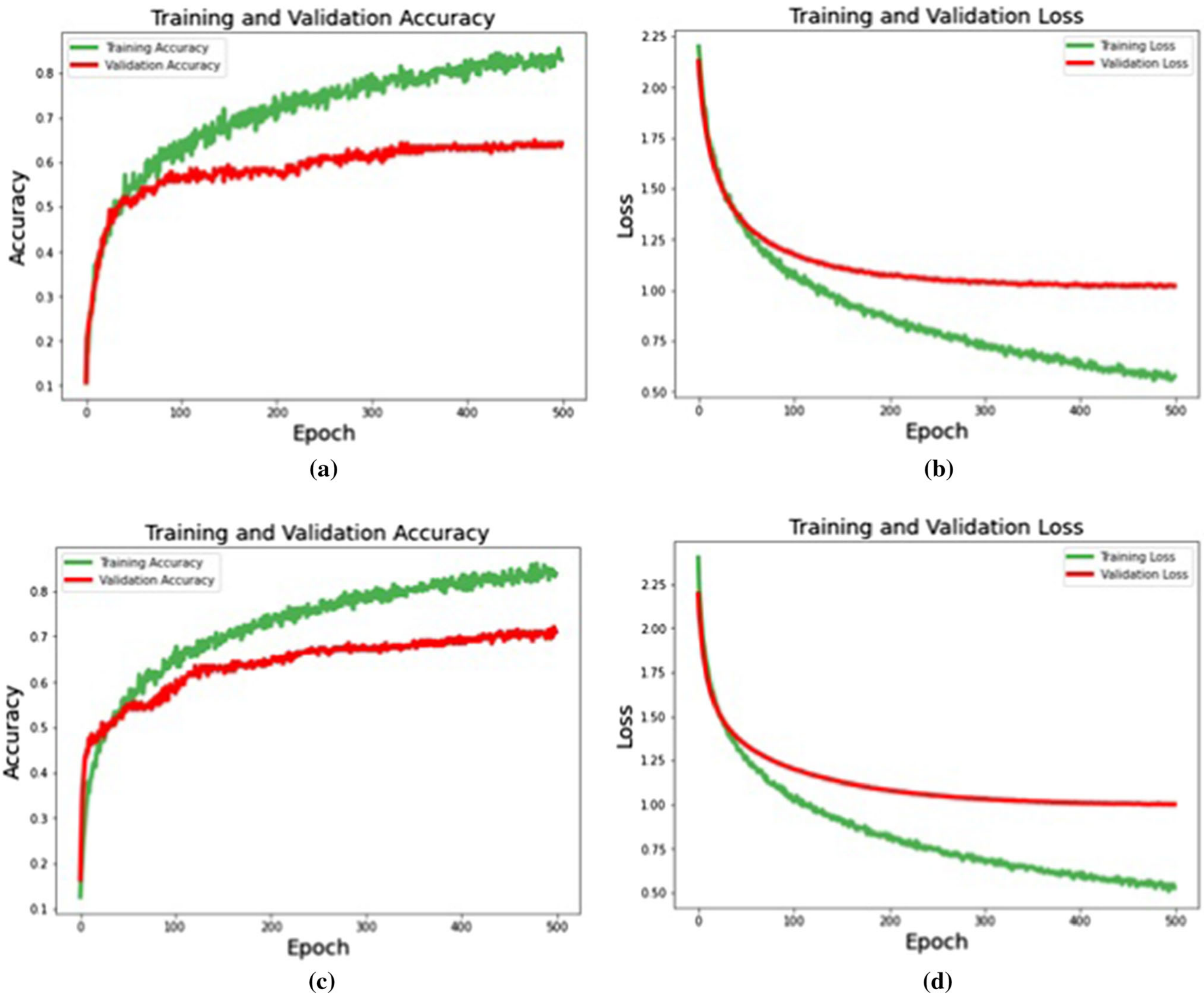
(a)

(b)

(c)

(d)

Fig. 7 Results using CNN model **a** TA and VA representation using Ballroom dataset, **b** TL and VL of representation using Ballroom dataset, **c** TA and VA representation using GTZAN dataset and **d** TL and VL of representation using GTZAN dataset



(a)

(b)

(c)

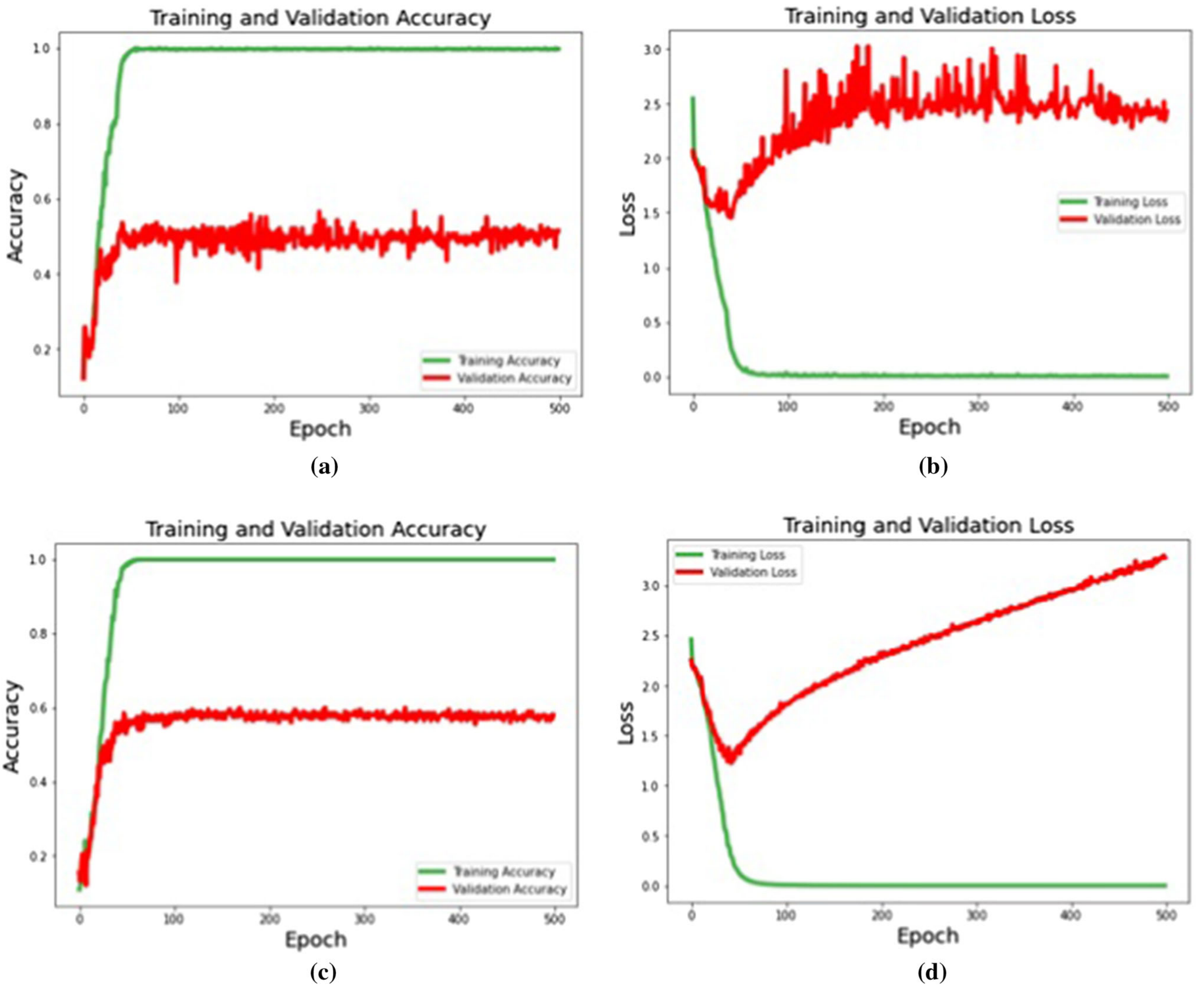
(d)

Fig. 8 Results using transfer learning-based model **a** TA and VA representation using Ballroom dataset, **b** TL and VL of representation using Ballroom dataset, **c** TA and VA representation using GTZAN dataset, and **d** TL and VL of representation using GTZAN dataset

Ballroom and GTZAN datasets, respectively. Figure 10b and d is focused on TL and VL representation by applying Ballroom and GTZAN datasets, respectively. In this work, Figs. 7, 8, 9 and 10 show the graphical representation of the change of TA, VA, TL and VL values by applying CNN, transfer learning, multimodal and proposed hybrid model, respectively, to Ballroom and GTZAN datasets.

In this work, the computed accuracy (percentage) as mentioned in Tables 1, 2, 3, 4, 5, 6, 7 and 8 is focused on the determination coefficient. Tables 1 and 2 describe the accuracy analysis of the CNN model using Ballroom and GTZAN datasets, respectively. From Table 1, it is observed that Waltz shows higher precision, recall and F1-score with values of 0.83, 0.86 and 0.84, respectively, and ChaChaCha shows higher support with a value of 23. The overall accuracy of this model is 59% using the Ballroom dataset. From Table 2, it is observed that Metal shows higher precision, recall and F1-score with values of 0.90 and the support value for each genre is 20. The overall accuracy of this model is 64% using the GTZAN dataset. From the results, it is observed that the CNN model is able to provide 59% accuracy using the Ballroom dataset and 64% accuracy using the GTZAN dataset.

For the CNN model using Ballroom dataset, it is found that the CM classifies correctly 14, 12, 11, 10, 6, 19, 7 and 4 music files of Tango, ChaChaCha, Rumba, Viennese waltz, Jlive, Waltz, Quickstep and Samba out of 17, 23, 20, 13, 12, 22, 16 and 17, respectively, which are represented diagonally. The genre Waltz has been classified well, whereas Samba is not being classified properly. From GTZAN datasets, it is found that the CM classifies correctly 11, 16, 8, 10, 14, 13, 18, 17, 9 and 11 music files out of 20, respectively, which are represented diagonally. The



(a)

(b)

(c)

(d)

Fig. 9 Results using multimodal model **a** TA and VA representation using Ballroom dataset, **b** TL and VL of representation using Ballroom dataset, **c** TA and VA representation using GTZAN dataset and **d** TL and VL of representation using GTZAN dataset

genre Metal has been classified well, whereas Country is not being classified properly.

Tables 3 and 4 describe the accuracy analysis of transfer learning-based model using Ballroom and GTZAN datasets, respectively. From Table 3, it is observed that Waltz and Tango show higher precision and recall with values of 0.75 and 0.94, respectively. Both Waltz and Tango show higher F1-score with values 0.78, and ChaChaCha shows higher support with a value of 23. The overall accuracy of this model is 64% using the Ballroom dataset. From Table 4, it is observed that Classical shows higher precision and F1-score with values 0.90, metal shows the recall value of 0.95 which is higher as compared to other genres, and each genre has a support value of 20. The overall accuracy of this model is 71% using GTZAN dataset. From the results, it is observed that the transfer learning-based model is able to provide 64% accuracy using Ballroom dataset and 71% accuracy using GTZAN dataset.

For transfer learning-based model using Ballroom dataset, it is found that the CM classifies correctly 16, 16, 7, 9, 9, 18, 11 and 4 music files of Tango, ChaChaCha, Rumba, Viennese waltz, Jlive, Waltz, Quickstep and Samba out of 17, 23, 20, 13, 12, 22, 16 and 17, respectively, which are represented diagonally. The genre Waltz has been classified well, whereas Samba is not being classified properly. From the GTZAN dataset, it is found that the CM classifies correctly 15, 18, 9, 14, 14, 17, 19, 15, 12 and 9 music files out of 20, respectively, which are represented diagonally. The genre metal has been classified well, whereas Country and Rock are not being classified properly.

Tables 5 and 6 describe the accuracy analysis of the multimodal model using Ballroom and GTZAN datasets,

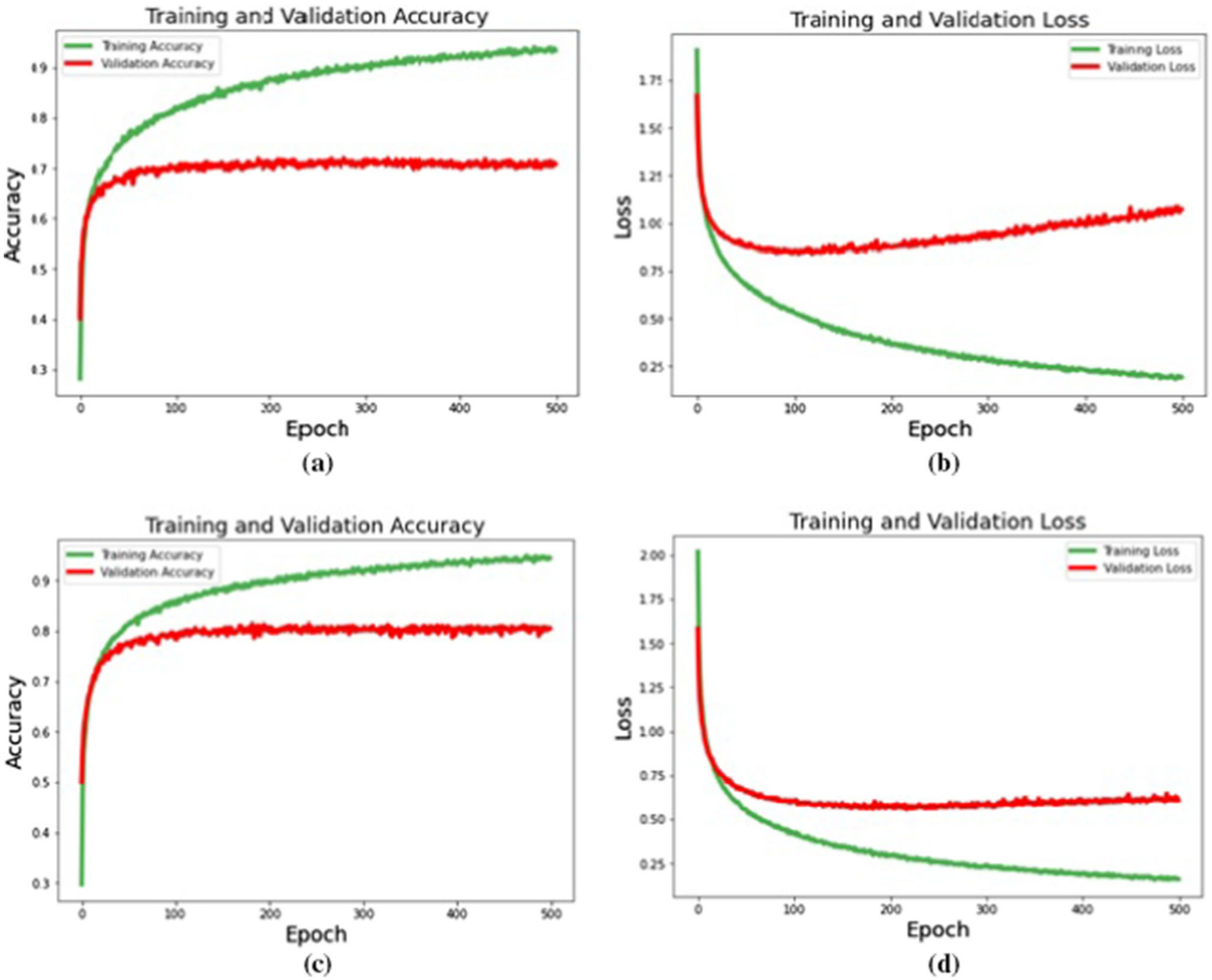


Fig. 10 Results using proposed hybrid model, **a** TA and VA representation using Ballroom dataset, **b** TL and VL of representation using Ballroom dataset, **c** TA and VA representation using GTZAN dataset, and **d** TL and VL of representation using GTZAN dataset

Table 1 Accuracy analysis of CNN model using Ballroom dataset

Genre	Precision	Recall	F1-score	Support
Tango	0.70	0.82	0.76	17
ChaChaCha	0.36	0.52	0.43	23
Rumba	0.58	0.55	0.56	20
Viennese waltz	0.77	0.77	0.77	13
Jlive	0.60	0.50	0.55	12
Waltz	0.83	0.86	0.84	22
Quickstep	0.47	0.44	0.45	16
Samba	0.57	0.24	0.33	17
Accuracy	59%			

Table 2 Accuracy analysis of CNN model using GTZAN dataset

Genre	Precision	Recall	F1-score	Support
Blues	0.58	0.55	0.56	20
Classical	0.80	0.80	0.80	20
Country	0.44	0.40	0.42	20
Disco	0.56	0.50	0.53	20
Hip-Hop	0.52	0.70	0.60	20
Jazz	0.72	0.65	0.68	20
Metal	0.90	0.90	0.90	20
Pop	0.89	0.85	0.87	20
Reggae	0.41	0.45	0.43	20
Rock	0.58	0.55	0.56	20
Accuracy	64%			

respectively. From Table 5, it is observed that Waltz shows higher precision, recall and F1-score with values 0.62, 0.91 and 0.74, respectively, and ChaChaCha shows a support value of 23 which is higher as compared to other genres.

Table 3 Accuracy analysis of transfer learning-based model using Ballroom dataset

	Precision	Recall	F1-score	Support
Tango	0.67	0.94	0.78	17
ChaChaCha	0.73	0.70	0.71	23
Rumba	0.58	0.35	0.44	20
Viennese waltz	0.69	0.69	0.69	13
Jlive	0.60	0.75	0.67	12
Waltz	0.75	0.82	0.78	22
Quickstep	0.50	0.69	0.58	16
Samba	0.50	0.24	0.32	17
Accuracy	64%			

Table 4 Accuracy analysis of transfer learning-based model using GTZAN dataset

	Precision	Recall	F1-score	Support
Blues	0.68	0.75	0.71	20
Classical	0.90	0.90	0.90	20
Country	0.69	0.45	0.55	20
Disco	0.56	0.70	0.62	20
Hip-Hop	0.67	0.70	0.68	20
Jazz	0.89	0.85	0.87	20
Metal	0.79	0.95	0.86	20
Pop	0.68	0.75	0.71	20
Reggae	0.57	0.60	0.59	20
Rock	0.69	0.45	0.55	20
Accuracy	71%			

Table 5 Accuracy analysis of multimodal model using GTZAN dataset

Genre	Precision	Recall	F1-score	Support
Tango	0.57	0.47	0.52	17
ChaChaCha	0.50	0.43	0.47	23
Rumba	0.36	0.25	0.29	20
Viennese waltz	0.40	0.31	0.35	13
Jlive	0.44	0.67	0.53	12
Waltz	0.62	0.91	0.74	22
Quickstep	0.42	0.31	0.36	16
Samba	0.60	0.71	0.65	17
Accuracy	51%			

Table 6 Accuracy analysis of multimodal model using GTZAN dataset

Genre	Precision	Recall	F1-score	Support
Blues	0.50	0.55	0.52	20
Classical	0.83	0.75	0.79	20
Country	0.50	0.40	0.44	20
Disco	0.60	0.60	0.60	20
Hip-Hop	0.41	0.60	0.49	20
Jazz	0.47	0.45	0.46	20
Metal	0.79	0.75	0.77	20
Pop	0.84	0.80	0.82	20
Reggae	0.52	0.55	0.54	20
Rock	0.41	0.35	0.38	20
Accuracy	58%			

Table 7 Accuracy analysis of proposed hybrid model using Ballroom dataset

Genre	Precision	Recall	F1-score	Support
Tango	0.73	0.75	0.74	170
ChaChaCha	0.78	0.71	0.74	230
Viennese waltz	0.52	0.39	0.45	130
Jlive	0.67	0.78	0.72	120
Waltz	0.73	0.81	0.77	220
Quickstep	0.61	0.71	0.65	160
Samba	0.76	0.74	0.75	170
Rumba	0.77	0.69	0.73	200
Accuracy	71%			

Table 8 Accuracy analysis of proposed hybrid model using GTZAN dataset

Genre	Precision	Recall	F1-score	Support
Country	0.72	0.70	0.71	200
Classical	0.94	0.95	0.95	200
Blues	0.82	0.73	0.75	200
Metal	0.90	0.91	0.90	200
Pop	0.79	0.82	0.82	200
Jazz	0.83	0.82	0.82	200
Rock	0.65	0.67	0.66	200
Disco	0.74	0.76	0.75	200
Hip-Hop	0.85	0.84	0.84	200
Raggae	0.83	0.81	0.82	200
Accuracy	81%			

The overall accuracy of this model is 51% using Ballroom dataset. From Table 6, it is observed that Pop shows higher precision, recall and F1-score with values 0.84, 0.80 and 0.82, respectively, and each genre has a support value of 20. The overall accuracy of this model is 58% using GTZAN dataset. From the results, it is observed that the multimodal model is able to provide 51% accuracy using Ballroom dataset and 58% accuracy using GTZAN dataset.

For the multimodal model using Ballroom dataset, it is found that the CM classifies correctly 8, 10, 5, 4, 8, 20, 5 and 12 music files of Tango, ChaChaCha, Rumba, Viennese waltz, Jlive, Waltz, Quickstep and Samba out of 17, 23, 20, 13, 12, 22, 16 and 17, respectively, which are represented diagonally. The genre Waltz has been classified well, whereas Viennese waltz is not being classified properly. From GTZAN datasets, it is found that the CM classifies correctly 11, 15, 8, 12, 12, 9, 15, 16, 11 and 7 music files out of 20, respectively, which are represented diagonally. The genre Pop has been classified well, whereas Rock is not being classified properly.

Tables 7 and 8 describe the accuracy analysis of the proposed hybrid model using Ballroom and GTZAN datasets, respectively. From Table 7, it is observed that ChaChaCha shows higher precision and support and its values are 0.78 and 230, respectively; Waltz shows higher recall and F1-score with values of 0.81 and 0.77, respectively. The overall accuracy of this model is 71% using Ballroom dataset. From Table 8, it is observed that Classical shows higher precision, recall and F1-score with values 0.94, 0.95 and 0.95, respectively, and each genre has a support value of 200. The overall accuracy of this model is 81% using GTZAN dataset. From the results, it is observed that the proposed hybrid model is able to provide 71% accuracy using Ballroom dataset and 81% accuracy using GTZAN dataset.

For the proposed hybrid model using Ballroom dataset, it is found that the CM classifies 13e + 02 music files of Tango correctly out of 170 which are represented in the diagonal with a darker color of the block. The CM classifies 16e + 02 music files of ChaChaCha correctly out of 230 which are represented in the diagonal with a dark color of the block. The CM classifies 51 music files of Viennese waltz correctly out of 130 which are represented in the diagonal with a lighter color of the block. The CM classifies 94 music files of Jlive correctly out of 120 which are represented in the diagonal with a lighter color of the block. The CM classifies 18e + 02 music files of Waltz correctly out of 220 which are represented in the diagonal with a very light color of the block. The CM classifies 11e + 02 music files of Quickstep correctly out of 160 which are represented in the diagonal with a darker color of the block. The CM classifies 13e + 02 music files of Samba correctly out of 170 which are represented in the diagonal with a very light color of the block. The CM classifies 14e + 02 music files of Rumba correctly out of 200 which are represented in the diagonal with a very light color of the block. This matrix also confuses Rumba with Waltz with a score of 14e + 02 of Rumba. Hence, the genre Jlive has been classified well, whereas Quickstep is not being classified properly.

From the GTZAN dataset, it is found that the CM classifies 15e + 02 music files of Blues correctly out of 20 which are represented in the diagonal with a darker color of the block. The CM classifies 19e + 02 music files of Classical correctly out of 20 which are represented in the diagonal with a darker color of the block. The CM classifies 14e + 02 music files of Country correctly out of 20 which are represented in the diagonal with a darker color of the block. The CM classifies 15e + 02 music files of Disco correctly out of 20 which are represented in the diagonal with a darker color of the block. The CM classifies 17e + 02 music files of Hip-Hop correctly out of 20 which are represented in the diagonal with a darker color of the block. The CM classifies 16e + 02 music files of Jazz correctly out of 20 which are represented in the diagonal with a darker color of the block. The CM classifies 18e + 02 music files of Metal correctly out of 20 which are represented in the diagonal with a very darker block. The CM classifies 17e + 02 music files of Pop correctly out of 20 which are represented in the diagonal with a darker color of the block. The CM classifies 16e + 02 music files of Raggae correctly out of 20 which are represented in the diagonal with a darker color of the block. The CM classifies 13e + 02 music files of Rock correctly out of 20 which are represented in the diagonal with a darker color of the block. This matrix also confuses Rock with Pop, Metal, Country and Blues with a score of 9,6,15 and 10, respectively. Hence, the genre Classical has been classified well, whereas Rock is not being classified properly.

From the analysis using Ballroom dataset, it is observed that Tango has the highest precision value of 0.73 in the proposed hybrid model and the lowest precision value of 0.57 in the multimodal model. ChaChaCha has the highest precision value of 0.78 in the hybrid model and the lowest precision value of 0.36 with CNN model. Rumba has the lowest precision value of 0.36 with the multimodal Model. Viennese waltz has the highest precision value of 0.77 in the CNN model and the lowest precision value of 0.40 in the multimodal model. Jlive has the highest precision value of 0.67 in the hybrid model and the lowest precision value of 0.44 in the multimodal model. Waltz has the highest precision value of 0.83 in the CNN Model and the lowest precision value of 0.62 in the multimodal model. Quickstep has the highest precision value of 0.61 in the hybrid model and the lowest precision value of 0.42 in the multimodal model. Samba has the highest precision value of 0.76 in the hybrid model and the lowest precision value of 0.50 in the transfer learning model. Tango has the highest recall value of 0.88 in the hybrid model and the lowest precision value of 0.47 in the multimodal model. ChaChaCha has the highest recall value of 0.71 in the hybrid model and the lowest precision value of 0.43 in the multimodal model. Rumba has the highest recall value of 0.69 in the hybrid model and the lowest precision value of 0.25 in the multimodal model. Viennese waltz has the highest recall value of 0.77 in the CNN model and the lowest precision value of 0.31 in the multimodal model. Jlive has the highest recall value of 0.78 in the hybrid model and the lowest precision value of 0.50 in the CNN model. Waltz has the highest recall value of 0.91 in the multimodel. Quickstep has the highest recall value of 0.71 in the hybrid model and the lowest precision value of 0.31 in the multimodal model. Samba has the highest recall value of 0.74 in the hybrid model and the lowest precision value of 0.24 in the transfer learning model and the CNN model both. Tango has the highest F1-score value of 0.78 in the transfer learning model and the lowest F1-score value of 0.52 in the multimodal model. ChaChaCha has the highest F1-score value of 0.74 in the hybrid model and the lowest F1-score value of 0.43 in the CNN model. Rumba has the highest F1-score value of 0.73 in the hybrid model and the lowest precision value of 0.29 in the multimodal model. Viennese waltz has the highest F1-score value of 0.77 in the CNN model and the lowest precision value of 0.35 in the multimodal model. Jlive has the highest F1-score value of 0.72 in the hybrid model and the lowest precision value of 0.53 in the multimodal model. Waltz has the highest F1-score value of 0.84 in the CNN model and the lowest precision value of 0.74 in the multimodal model. Quickstep has the highest F1-score value of 0.65 in the hybrid model and the lowest precision value of 0.36 in the multimodal model. Samba has the highest F1-score value of 0.75 in the proposed hybrid model and the lowest precision value of 0.32 in the transfer learning model.

From the analysis using GTZAN dataset, it is observed that Blues has the highest precision value of 0.82 in the hybrid model and the lowest precision value of 0.50 in the multimodal model. Classical has the highest precision value of 0.94 in the hybrid model and the lowest precision value of 0.80 in the CNN model. Country has the highest precision value of 0.83 in the hybrid model and the lowest precision value of 0.44 with the CNN model. Disco has the highest precision value of 0.74 in the hybrid model and the lowest precision value of 0.55 in the hybrid model. Hip-hop has the highest precision value of 0.85 in the hybrid model and the lowest precision value of 0.41 with the multimodal model. Jazz has the highest precision value of 0.89 in the transfer learning model and the lowest precision value of 0.47 in the multimodal model. Metal has the highest precision value of 0.90 in the hybrid model and the CNN model and the lowest precision value of 0.79 in the proposed hybrid model. Pop has the highest precision value of 0.89 in the CNN model and the lowest precision value of 0.68 in the transfer learning model. Raggae has the highest precision value of 0.83 in the hybrid model and the lowest precision value of 0.41 in the CNN model. Rock has the highest precision value of 0.69 in the transfer learning model and the lowest precision value of 0.41 in the multimodal model. Blues has the highest recall value of 0.75 in both the proposed hybrid and the transfer learning model and the lowest precision value of 0.55 in both the CNN and the multimodal model. Classical has the highest recall value of 0.95 in the proposed hybrid model and the lowest precision value of 0.75 in the multimodal model. Country has the highest recall value of 0.70 in the hybrid model and the lowest precision value of 0.40 for both the CNN and the multimodal model. Disco has the highest recall value of 0.80 in the hybrid model and the lowest precision value of 0.50 in the CNN model. Hip-hop has the highest recall value of 0.84 in the hybrid model and the lowest precision value of 0.60 in the multimodal model. Jazz has the highest recall value of 0.85 in both the transfer learning and the hybrid model and the lowest precision value of 0.45 in the multimodal model. Metal has the highest recall value of 0.95 in the transfer learning model and the lowest precision value of 0.75 in the multimodal model. Pop has the highest recall value of 0.85 in the CNN model and the lowest precision value of 0.75 in both the transfer learning model and the proposed hybrid model. Raggae has the highest recall value of 0.81 in the hybrid model and the lowest precision value of 0.45 in the CNN model. Rock has the highest recall value of 0.67 in the hybrid model and the lowest precision value of 0.35 in the multimodal model. Blues has the highest F1-score value of 0.75 in the hybrid model and the lowest precision value of 0.52 in the multimodal model. Classical has the highest F1-score value of 0.95 in the hybrid model and the lowest precision value of 0.79 in the multimodal model. Country has the highest F1-score value of 0.71 in the hybrid model and the lowest precision value of 0.42 with the CNN model. Disco has the highest F1-score value of 0.75 in the hybrid model and the lowest precision value of 0.53 in the CNN model. Hip-hop has the highest F1-score value of 0.84 in the hybrid model and the lowest precision value of 0.49 in the multimodal model. Jazz has the highest F1-score value of 0.87 in the transfer learning-based model and the lowest precision value of 0.46 in the multimodal model. Metal has the highest F1-score value of 0.90 in both the hybrid model and the CNN model and the lowest precision value of 0.77 in the multimodal model. Pop has the highest F1-score value of 0.87 in the CNN model and the lowest precision value of 0.71 in the transfer learning-based model. Raggae has the highest F1-score value of 0.82 in the hybrid model and the lowest precision value of 0.43 in the CNN model. Rock has the highest F1-score value of 0.66 in the hybrid model and the lowest precision value of 0.38 in the proposed hybrid Model. The overall performance of the proposed hybrid model with deep learning models, machine learning models (SVM and NN) and other existing models using Ballroom and GTZAN datasets are mentioned in Table 9, and graphically, it is represented using Fig. 11.

From the above analysis, it is observed that the proposed hybrid model is able to classify the music genre files available in the Ballroom dataset and GTZAN dataset in a better way as compared to the other models. The proposed hybrid model is able to provide improvement of 10%, 16%, 8%, 10%, 9%, 9%, 18%, 0%, 22%, 9%, 71%, 71%, 17%, 10% and 23% more than JSLRR [21], JSR [21], LRR [21], SRC [21], LRC [21], SVM [21], NN [21], baseline [22], rhythmic [22], combined [22], *P*.-MaxP [22], *P*.-Avg [22], CNN, transfer learning and multimodal learning, respectively, when the dataset is GTZAN. When the dataset is Ballroom, the proposed hybrid model is able to provide improvement of 10%, 15%, 9%, 12%, 12%, 9%, 21%, 11%, 4%, 0%, 56%, 59%, 12%, 7% and 20% more than JSLRR [21], JSR [21], LRR [21], SRC [21], LRC [21], SVM [21], NN [21], baseline [22], rhythmic [22], combined [22], P.-MaxP [22], P.-Avg [22], CNN, transfer learning, and multimodal learning, respectively.

Here, one-tail paired t-test is performed using Python to show the variance of the accuracy of the models in two different datasets as per Table 9. The mean of accuracy of all models for GTZAN dataset and Ballroom dataset is found to be 61.5 and 54.937. The variance is found to be 437.866 and 295.662, respectively. The standard deviation is found to be 20.925 and 17.195, respectively. The confidence interval at 95% is [51.247, 71.753] and [46.512, 63.362], respectively. The T-score is found to be 0.9692, degree of freedom is 30, standard error of difference is 6.771, and P-value is 0.3402. From the result, the null hypothesis is rejected at a P-value of 0.05 and this difference is considered to be not statistically significant or variance is there between the accuracies of the models of two different datasets with the confidence of 95%.

The computational time analysis is shown in Table 10 for the proposed hybrid model. Here, the experiments are performed on two machines. The first one is laptop, and second is PARAM SHAVAK supercomputer. The laptop has the following configuration such as OS—Win 10 × 64, CPU—AMD A9-9420 RADEON R5, 5 Compute Cores (2C + 3G)–3.0 GHz, RAM—8 GB, 1066 MHz and HDD—1 TB. The PARAM SHAVAK supercomputer has the following configuration such as OS—Ubuntu Linux, 28

Table 9 Accuracy (%) of the proposed hybrid model with deep learning models, machine learning models and other existing models

Model	GTZAN (%)	Percentage (%) of improvement of proposed hybrid model using GTZAN dataset	Ballroom (%)	Percentage (%) of improvement of proposed hybrid model using Ballroom dataset
JSLRR [21]	71	10	61	10
JSR [21]	65	16	56	15
LRR [21]	73	8	62	9
SRC [21]	71	10	59	12
LRC [21]	72	9	59	12
SVM [21]	72	9	62	9
NN [21]	63	18	50	21
Baseline [22]	72	9	60	11
Rhythmic [22]	59	22	67	4
Combined [22]	72	9	71	0
P-MaxP [22]	10	71	15	56
P-Avg [22]	10	71	12	59
CNN	64	17	59	12
Transfer Learning	71	10	64	7
Multimodal	58	23	51	20
Proposed Hybrid model	81	–	71	–

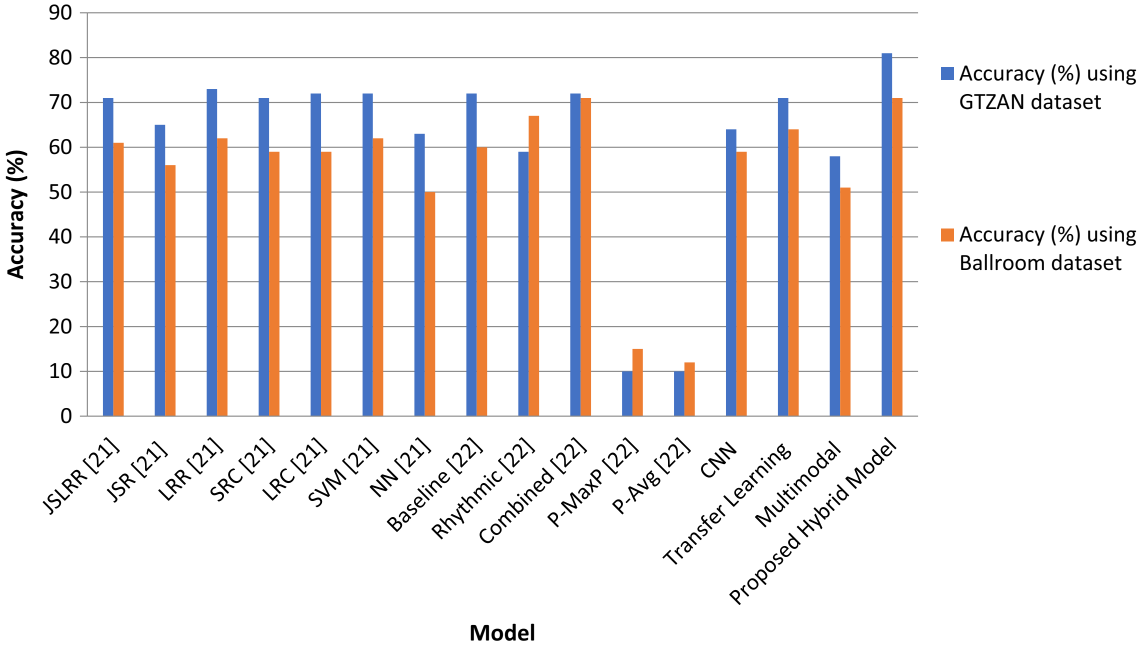


Fig. 11 Accuracy (%) representation of the proposed hybrid model and other models

Table 10 Computation time analysis of proposed hybrid model

Dataset	Number of epoch	Total steps	Execution time in laptop (in second)			Execution time in PARAM SAVAK (in second)		
			Per step	Per epoch	Total	Per step	Per epoch	Total
GTZAN	500	25	2.52	63	31578 s (8.77 Hr.)	0.64	16	8000 s (2.22 Hr.)
Ballroom	500	18	3.64	65.52	32,760 (9.1 Hr.)	0.94	17	8500 s (2.36 Hr.)

core processor, 96 GB RAM, 25 TB HDD, and NVIDIA GPU 8 GB. The computational results are discussed below as follows.

1. Computation time in laptop: From the result, it is seen that when the dataset is GTZAN with epoch = 500, steps = 25, then the execution time in the laptop is shown to be 2.52 s per epoch, 63 s per epoch, and total time is 31578 s or 8.77 Hr. When the dataset is Ballroom with epoch = 500, steps = 18, then the execution time in laptop is shown to be 3.64 s per epoch, 65.52 s per epoch, and total time is 32760 s or 9.1 Hr.
2. Computation time in PARAM SHAVAK supercomputer: When the dataset is GTZAN with epoch = 500, steps = 25, then the execution time in the laptop is shown to be 0.64 s per epoch, 16 s per epoch, and the total time is 8000 s or 2.22 Hr. When the dataset is Ballroom with epoch = 500, steps = 18, then the execution time in the laptop is shown to be 0.94 s per epoch, 17 s per epoch, and total time is 8500 s or 2.36 Hr.

5 Conclusion

This work proposed a DL-based approach for the analysis and classification of music genre files. The DL-based approach is focused on the models such as the CNN model, transfer learning-based model, multimodal training model, and the proposed hybrid model to carry out such classification. These models are analyzed using the GTZAN and Ballroom datasets. The performance of these models is evaluated using the performance parameters such as training accuracy, validation accuracy, training loss, validation loss, precision, recall, F1-score and support. The macro-average and weighted average are taken for computing the percentage of accuracy of each model. From the results, it is concluded that the proposed hybrid model performs better with 81% and 71% accuracy using GTZAN and Ballroom datasets, respectively, as compared to other models. The proposed hybrid model is able to show improvement of 10%, 16%, 8%, 10%, 9%, 9%, 18%, 0%, 22%, 9%, 71%, 71%, 17%, 10% and 23% more than JSLRR [21], JSR [21], LRR [21], SRC [21], LRC [21], SVM [21], NN [21], baseline [22], rhythmic [22], combined [22], P.-MaxP [22], P.-Avg [22], CNN, transfer learning and multimodal learning, respectively, when the dataset is GTZAN. When the dataset is Ballroom, proposed hybrid model shows improvement of 10%, 15%, 9%, 12%, 12%, 9%, 21%, 11%, 4%, 0%, 56%, 59%, 12%, 7% and 20% more than JSLRR [21], JSR [21], LRR [21], SRC [21], LRC [21], SVM [21], NN [21], baseline [22], rhythmic [22], combined [22], *P*.-MaxP [22], *P*.-Avg [22], CNN, transfer learning and multimodal learning, respectively. This work can be extended to develop enhanced methods to improve the accuracy percentage on the same datasets as well as some other datasets. New deep learning models can also be considered for such classification for improving the accuracy and reducing error.

Acknowledgements We want to thank Department of CSE, PMEC Berhampur (Government), India for providing adequate infrastructure and facilities to conduct this research work. We also want to thank Sonalisha Mohapatra for her overall coordination with group members Anup Pradhan, Subhankar Dash, and Subham Kumar Biswal of Department of CSE, PMEC Berhampur for supporting in completion of this research work under guidance of Dr. Kalyan Kumar Jena and Dr. Sourav Kumar Bhoi.

Author contributions KKJ, SKB and SM contributed equally to this whole work. SB contributed in the design of hybrid model and simulations.

Funding There is no funding information.

Data availability Data are available on request.

Declarations

Conflict of interest There is no conflict of interest.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent to publication Not applicable.

References

1. Oramas S, Barbieri F, Nieto Caballero O, Serra X (2018) The Multimodal deep learning for music genre classification. Trans Int Soc Music Inf Retr 1(1):4–21. https://doi.org/10.5334/tismir.10
2. Feng T (2014) Deep learning for music genre classification. Private document. pp. 1–7. https://courses.engr.illinois.edu/ece544na/fa2014/Tao_Feng.pdf
3. Bahuleyan H (2018) Music genre classification using machine learning techniques. arXiv preprint arXiv:1804.01149
4. Elbir A, Aydin N (2020) Music genre classification and music recommendation by using deep learning. Electron Lett 56(12):627–629. https://doi.org/10.1049/el.2019.4202
5. Nanni L, Costa YM, Aguiar RL, Silla CN Jr, Brahnam S (2018) Ensemble of deep learning, visual and acoustic features for music genre classification. J New Music Res 47(4):383–397. https://doi.org/10.1080/09298215.2018.1438476
6. Kim S, Kim D Suh B (2016) Music genre classification using the multimodal deep learning. In: Proceedings of HCI Korea pp. 389–395. https://doi.org/10.17210/hcik.2016.01.389
7. Oramas S, Nieto O, Barbieri F, Serra X (2017) Multi-label music genre classification from audio, text, and images using deep features. arXiv preprint arXiv:1707.04916
8. Vishnupriya S, Meenakshi K (2018) Automatic music genre classification using convolution neural network. In: 2018 International conference on computer communication and informatics (ICCCI). IEEE pp. 1–4. https://doi.org/10.1109/ICCCI.2018.8441340
9. Lau DS, Ajoodha R (2022) Music genre classification: a comparative study between deep learning and traditional machine learning approaches. In: Proceedings of sixth international congress on information and communication technology. Springer, Singapore pp. 239–247. https://doi.org/10.1007/978-981-16-2102-4_22
10. Jeong IY, Lee K (2016) Learning temporal features using a deep neural network and its application to music genre classification. In: Ismir pp. 434–440. https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/159_Paper.pdf
11. Senac C, Pellegrini T, Mouret F, Pinquier J (2017) Music feature maps with convolutional neural networks for music genre classification. In: Proceedings of the 15th international workshop on content-based multimedia indexing pp. 1–5. https://doi.org/10.1145/3095713.3095733
12. Yu Y, Luo S, Liu S, Qiao H, Liu Y, Feng L (2020) Deep attention based music genre classification. Neurocomputing 372:84–91. https://doi.org/10.1016/j.neucom.2019.09.054
13. Aguiar RL, Costa YM, Silla CN (2018) Exploring data augmentation to improve music genre classification with convnets. In: 2018 International joint conference on neural networks (IJTHE CNN), IEEE pp. 1–8. https://doi.org/10.1109/IJCNN.2018.8489166
14. Yang R, Feng L, Wang H, Yao J, Luo S (2020) Parallel recurrent convolutional neural networks-based music genre classification method for mobile devices. IEEE Access 8:19629–19637. https://doi.org/10.1109/ACCESS.2020.2968170
15. Zhang W, Lei W, Xu X, Xing X (2016) Improved music genre classification with convolutional neural networks. In: Interspeech pp. 3304–3308. https://www.isca-speech.org/archive_v0/Interspeech_2016/pdfs/1236.PDF
16. Liu J, Wang C, Zha L (2021) A middle-level learning feature interaction method with deep learning for multi-feature music genre classification. Electronics 10(18):2206. https://doi.org/10.3390/electronics10182206
17. Rajanna AR, Aryafar K, Shokoufandeh A, Ptucha R (2015) Deep neural networks: a case study for music genre classification. In: 2015 IEEE 14th international conference on machine learning and applications (ICMLA), IEEE pp. 655–660. https://doi.org/10.1109/ICMLA.2015.160
18. Shi L, Li C, Tian L (2019) Music genre classification based on chroma features and deep learning. In: 2019 Tenth international conference on intelligent control and information processing (ICICIP), IEEE pp. 81–86. https://doi.org/10.1109/ICICIP47338.2019.9012215

19. Elbir A, Çam HB, Iyican ME, Öztürk B, Aydin N (2018). Music genre classification and recommendation by using machine learning techniques. In: 2018 Innovations in intelligent systems and applications conference (ASYU), IEEE pp. 1–5. https://doi.org/10.1109/ASYU.2018.8554016
20. Tsaptsinos A (2017) Lyrics-based music genre classification using a hierarchical attention network. arXiv preprint arXiv:1707.04678
21. Panagakis Y, Kotropoulos CL, Arce GR (2014) Music genre classification via joint sparse low-rank representation of audio features. IEEE/ACM Trans Audio Speech Lang Process 22(12):1905–1917. https://doi.org/10.1109/TASLP.2014.2355774
22. Lykartsis A, Lerch A (2015) Beat histogram features for rhythm-based musical genre classification using multiple novelty functions. In: 18th International conference on digital audio effects. Trondheim, Norway, pp.1–8. https://musicinformatics.gatech.edu/wp-content_nondefault/uploads/2015/12/DAFx-15_submission_42-1.pdf
23. http://mtg.upf.edu/ismir2004/contest/tempoContest/node5.html, accessed on Sep 2021
24. https://www.kaggle.com/andradaolteanu/gtzan-dataset-music-genre-classification, accessed on Sep 2021
25. 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), IEEE pp. 974–978. https://doi.org/10.1109/ICCMC53470.2022.9753953
26. Hongdan W, SalmiJamali S, Zhengping C, Qiaojuan S, Le R (2022) An intelligent music genre analysis using feature extraction and classification using deep learning techniques. Comput Elect Eng 100:107978. https://doi.org/10.1016/j.compeleceng.2022.107978
27. Falola PB, Alabi EO, Ogunajo FT, Fasae OD (2022) Music genre classification using machine and deep learning techniques: a review. ResearchJet J Anal Invent 3(03):35–50
28. Singh Y, Biswas A (2022) Robustness of musical features on deep learning models for music genre classification. Expert Syst Appl 199:116879. https://doi.org/10.1016/j.eswa.2022.116879
29. Wang W, Sohail M (2022) Research on music style classification based on deep learning. Comput Math Methods Med 2022:1–8. https://doi.org/10.1155/2022/3699885
30. Narkhede, N., Mathur, S., & Bhaskar, A. (2022). Machine learning techniques for music genre classification. In: Information and communication technology for competitive strategies (ICTCS 2020). Springer, Singapore pp. 155–161. https://doi.org/10.1007/978-981-16-0739-4_15
31. Gupta R, Ashish S, Shekhar H, Dominic MS (2022) Music genre classification using CNN and RNN-LSTM. In: Micro-electronics and telecommunication engineering. Springer, Singapore

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.