Hindi Songs Genre Classification Using Deep Learning

Md Shah Fahad, Raushan Raj, Ashish Ranjan, and Akshay Deepak

1 Introduction

Music is like a mirror; it provides a lot of information about who you are and what you like. Companies increasingly utilize music classification to classify what they care about either to be able to make consumer recommendations (such as Spotify, Soundcloud, and other similar services) or just as a product (for example, Shazam) [1]. The first step in that direction is to identify music genres. Machine Learning algorithms are effective at extracting trends and patterns from massive amounts of data. In music analysis, the same principles are used [2, 3]. Songs are analyzed for tempo, acoustics, energy, and other factors based on their digital signatures.

Early notable works for the music genre classification include [1, 2, 4, 5]. In the paper [4], a multivariate auto-regressive feature model is introduced for music genre classification. In [5], the author proposed an ensemble method based on the deep learning (extracted using CNN) + hand-crafted features (time-domain features and frequency-domain features) for the seven-class music genre classification. In [1], a (CNN + word2vec) model is used for predicting the music genre. Most of these approaches use MFCC as the input feature for their deep learning models [1, 5]. In the paper [6], the authors proposed the method for classifying Hindi songs into four genre classes—*Classical*, *Folk*, *Ghazal*, and *Sufi*. They employed spectral features and SVM classifiers to conduct the classification. Further, in [7], the author studied and identified hand-crafted features specific to music genre classification that include mel-frequency cepstrum coefficients (MFCCs), pitch [8], dominant frequency, and chroma [9] features.

M. S. Fahad (✉) · R. Raj · A. Ranjan · A. Deepak
National Institute of Technology, Patna, India
e-mail: shah.cse16@nitp.ac.in

R. Raj
e-mail: raushanr.pg20.cs@nitp.ac.in

A. Ranjan
e-mail: ashish.cse16@nitp.ac.in

A. Deepak
e-mail: akshayd@nitp.ac.in

U. P. Rao et al. (eds.), *Security, Privacy and Data Analytics*, Lecture Notes in Electrical Engineering 848, https://doi.org/10.1007/978-981-16-9089-1_12

The proposed framework used a convolutional neural network (CNN) alongside the long short-term memory network (LSTM) with an attention mechanism for the task of Hindi music genre classification. In this work, a convolutional neural network (CNN) [10] is used to determine the features automatically from data itself [5]. Chroma, Pitch, Mel-spectrogram [11], and MFCC are used as input to the CNN. Further, the global features are identified by using a long short-term memory network (LSTM). LSTM [12] can learn the sequential pattern of different categories of songs. Further, because each sub-part of a signal does not contribute equally, henceforth, an attention mechanism [13] is applied to weigh each sub-part. The results for the proposed framework demonstrate the effectiveness of the MFCC features for the audio genre classification.

Paper organization: The explanation of the dataset and the proposed methodology are discussed in Sect. 2. Section 3 discusses the experimental design and results. Section 4 concludes the paper.

2 Proposed Methodology

Further, this section is divided into subsections as demonstrated below. Section 2.1 describes the database. Section 2.2 discusses feature extraction. The discussion of deep learning architecture is presented in Sect. 2.3.

2.1 Dataset

The dataset is composed of 599 audio tracks of Hindi songs where each audio track is trimmed to a maximum length of 30 s. Further, each audio track is trimmed between 0.5 and 5.5 s to generate a feasible size of the input matrix. The sampling rate is 40 kHz. The genres corresponding to songs were downloaded manually taking references from popular websites, such as Spotify and Wynk. It includes six genres: Sufi, Classical, Romantic, Ghazal, Party, and Bhakti. Each genre has between 70 and 80 sound clips. The dataset is collected from various websites in which the categorization of songs is given. The distribution of samples corresponding to each genre is shown in Fig. 1.

2.2 Feature Representation

Every audio signal contains a variety of features. However, we must extract the features that are pertinent to the problem at hand. Librosa [14] is a Python module for analyzing audio signals in general, with a focus on music. In this work, chroma, pitch, mel-spectrogram, and MFCC features are used as input for the proposed framework.

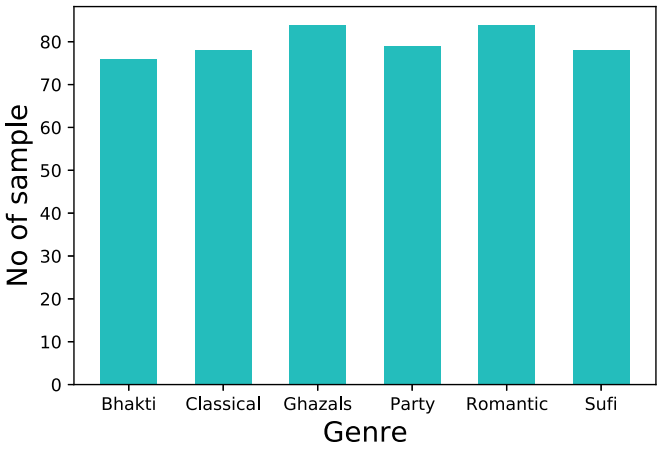

Fig. 1 Distribution of a number of audio signals of each song's genre

The audio signal is non-stationary, therefore, frequency transformation is done frame-wise with 50% overlapping. The size of a frame is the same as FFT length, i.e. 2048. The total number of frames is 391 for an audio clip. The other attributes for different representations are:

1. **Chroma**: For chroma features, 12 bins are used and an output matrix of size 431×12 is produced.
2. **Pitch**: For pitch, 1025 FFT bins are chosen, resulting into 1025 pitch and their corresponding magnitude values for each frame. A matrix of size $391 \times 1025 \times 2$ is created. [1]
3. **Mel-spectrogram** [11]: The mel-frequency cepstrum captures the properties of the signal's frequency as represented on the Mel-scale, which closely resembles the non-linearity of human hearing. For Mel-spectrogram, the frequency of a signal is represented on the Mel-scale of length 128 which is similar to the non-linear nature of the human hearing. Thus, a matrix of size 391×128 is fed to the proposed framework. Mel-spectrogram corresponding to each of the song's genre is shown in Fig. 2.

4. **MFCC**: MFCCs are often the frequently used features for several speech-related tasks [15]. For MFCC, 26 MFCC features are extracted and a matrix of size 391×26 is fed to the proposed framework. MFCC corresponding to each of the song's genres is shown in Fig. 3.

[1] https://librosa.org/doc/main/generated/librosa.piptrack.html

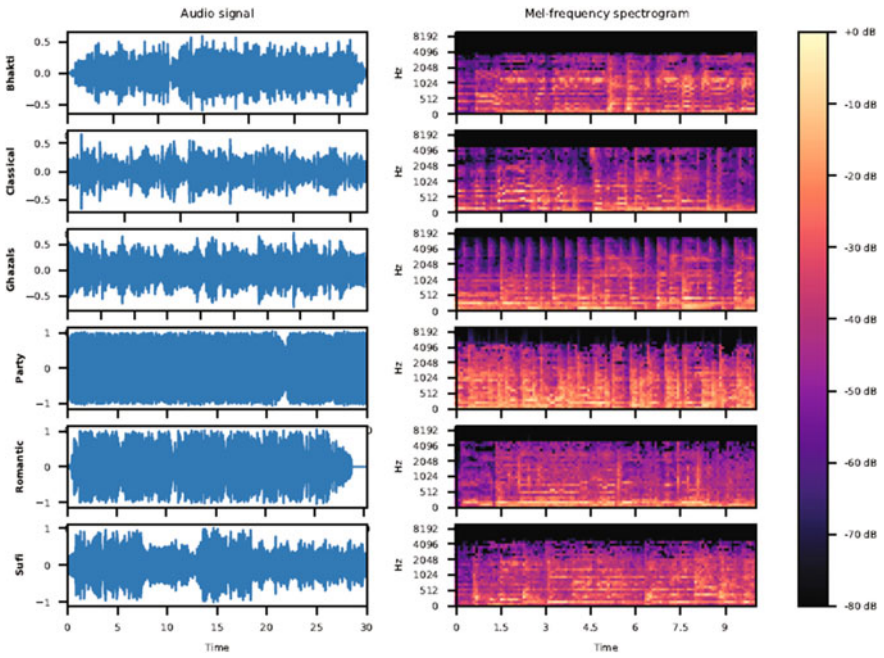

Fig. 2 Audio signal and corresponding Mel-Spectrogram of each song's genre

2.3 Proposed Framework

The proposed solution is a deep learning-based framework and consists of CNN, bi-directional LSTM, and Attention [13] layers. In contrast to shallow neural networks, deep neural networks are more specialized toward extracting the features that are more meaningful and incorporate a better learning methodology. The block diagram of the proposed framework is depicted in Fig. 3. These layers are described next:

1. **CNN**: The advantage of convolutional neural network (CNN) [10] layers is that they aid in capturing the local features by the convolution operation.

   A stack of two (convolution + batch-normalization) layers is used. The input to the first convolutional layer is a matrix representation of an audio clip. The matrix can either be a representation, for instance, Pitch, Chroma, Mel-spectrogram, and MFCC. After that, max-pooling layers are added to minimize the number of parameters. The hyper-parameters with CNN layers are as follows:

   (a) Filter-size (1st layer): 5×5
   (b) Filter-size (2nd layer): 5×5
   (c) Number of filters (1st layer): 32
   (d) Number of filters (2nd layer): 64

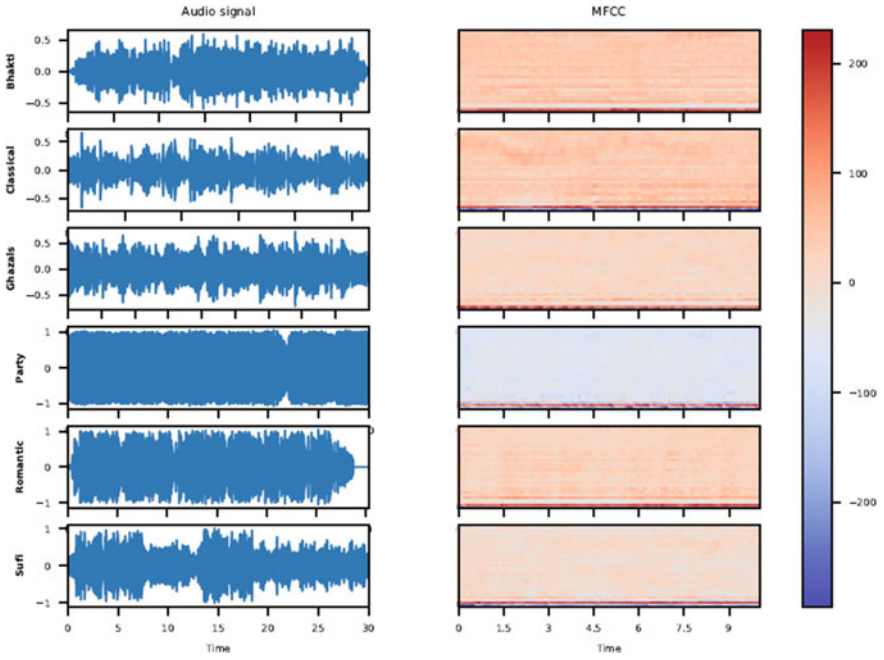

Fig. 3 Audio signal and corresponding MFCC of each song's genre

(e) Pool-size: (2×2).

2. **LSTM**: LSTM is popular for sequences, such as NLP, biological sequences [16], and speech signals [12, 17] and aid in learning temporal dependencies in the sequence.

   The sequential behavior underlying the features extracted from the CNN layer is learned by using the long short-term memory (LSTM) layers. The LSTM network is used in a bi-directional mode [12] and the number of hidden neuron units with the LSTM cell is taken as 32.

3. **Attention**: This helps assign different weights to the LSTM outputs, since not all the regions in a given signal play the same role. The additive attention proposed by Bahdanau et al. [13] was used for this work.

In between the CNN layer and LSTM layer is the reshape layer to transform the 3-D output to a 2-D output (suitable for the LSTM).

Further, a dense layer is stacked on top of the attention layer. The number of neurons in this layer is 20, while the activation function is taken as *ReLU*. Finally, the last layer with *Softmax* activation is used to perform the music genre classification task.

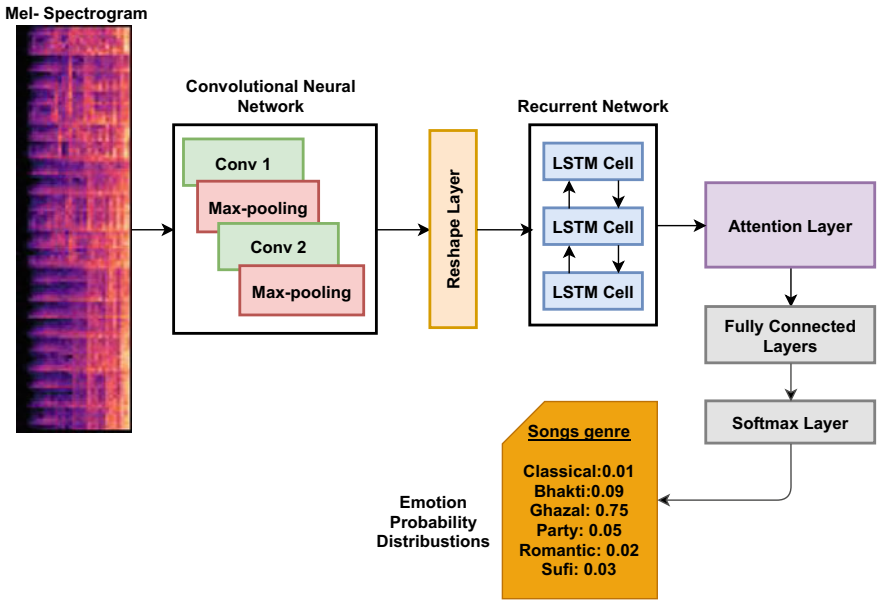

Fig. 4 Block diagram of the proposed framework

The choice of other hyper-parameters for the proposed framework is as follows (Fig. 4):

1. Loss function: categorical_cross-entropy to deal multi-class problem.
2. Optimizer: Adam.
3. Batch-size: 32.
4. Total epochs: 100.
5. Early-stopping criteria with patience: 10 epoch.

3 Experimental Design and Result Discussion

The proposed model architecture was implemented using the Keras API that is built on top of TensorFlow. The dataset is divided into training and testing sets, with 80% data (479 samples) in the training set and 20% data (120 samples) in the testing set. The training set is used for developing the model and the test set for evaluating the model.

Table 1 Precision (P), Recall (R), and F1-Score (F1) for each feature (pitch, chroma, Mel-spectrogram (Mel), and MFCC) corresponding to each genre

Feature	Metric	Bhakti	Classical	Ghazal	Party	Romantic	Sufi	Avg.
Pitch	P	0.45	0.41	0.72	0.65	0.33	0.33	0.48
	R	0.42	0.50	0.81	0.71	0.21	0.33	0.49
	F1	0.43	0.45	0.76	0.68	0.26	0.33	0.48
Chroma	P	0.32	0.41	0.62	0.58	0.32	0.44	0.44
	R	0.33	0.50	0.62	0.67	0.32	0.22	0.44
	F1	0.33	0.45	0.62	0.62	0.32	0.30	0.43
Mel	P	0.50	0.59	0.93	0.70	0.50	0.38	0.59
	R	0.58	0.59	0.88	0.76	0.42	0.33	0.59
	F1	0.54	0.59	0.90	0.73	0.46	0.35	0.59
MFCC	P	**0.70**	**0.73**	**1.00**	**0.82**	**0.52**	**0.62**	**0.73**
	R	**0.67**	**1.00**	**0.88**	**0.67**	**0.63**	**0.44**	**0.71**
	F1	**0.68**	**0.85**	**0.93**	**0.74**	**0.57**	**0.52**	**0.71**

3.1 Comparison with Respect to Different Speech Features

In this work, different speech features, namely (i) Chroma, (ii) Pitch, (iii) Mel-spectrogram, and (iv) Mel Frequency Cepstral Coefficient (MFCC) have been explored using the proposed deep learning (CNN + Bi-directional LSTM + Attention) architecture. Popular evaluation metrics such as *Precision*, *Recall*, and *F1-score* for each speech features (Pitch, Chroma, Mel-spectrogram, and MFCC) for different genre classes (Bhakti, Classical, Ghazal, Party, Romantic, and Sufi) are shown in Table 1.

The best precision, recall, and F1-score are obtained using MFCC features. The genre class "*Ghazal*" has the best precision, recall, and F1-score 100%, 88%, and 93%, respectively. After that, the genre classes "*Classical*", "*Party*", and "*Bhakti*" achieve better results order-wise. The genre class "*Sufi*" is least accurate among all the genres. It achieves only 52% F1-score. The genre class "*Sufi*" is more confused with the genre "*Bhakti*" and "*Romantic*".

3.2 Comparison with Other State-of-the-Art Models

This includes a comparison to the other standard deep learning models, such as the standalone CNN model and the (CNN + LSTM) model, widely applicable in processing the speech signal data. Note that the LSTM model is evaluated for both; the uni-directional and the bi-directional configuration. This sort of comparison also benefits understanding the advantage of different layers with the proposed model (i.e. based on the MFCC features). The results obtained for the different models are shown in Table 2.

Table 2 Classification report with respect to different models (Uni: stands for the uni-direction LSTM; Bi: stands for the bi-direction LSTM; *indicated the proposed model)

S. no.	Models	Precision	Recall	F1-Score
1.	CNN	0.57	0.57	0.56
2.	CNN + Uni-LSTM	0.59	0.62	0.58
3.	CNN + Bi-LSTM	0.62	0.65	0.63
4.	CNN + Uni-LSTM + Attention	0.68	0.68	0.67
5.	CNN + Bi-LSTM +Attention*	**0.73**	**0.71**	**0.71**

As observed from Table 2, the results clearly suggest the strong results for the proposed model. The proposed model beats the standalone CNN-based model by a significantly huge margin, where the improvements recorded with respect to the *F1-score* is 15.0 percent. Similarly, the proposed model also outperformed the (CNN + Bi-LSTM) model by a margin of 8.0 percent. These improvements were also observed for other metrics, such as *precision* and *recall*. In addition, performances obtained for models with the uni-directional LSTM come out to be much lower when they are compared to models with the bi-directional LSTM.

The confusion matrices corresponding to different experimental models are also shown in Fig. 5 (with the CNN model), Fig. 6 (with the CNN + Uni-directional LSTM), Fig. 7 (with the CNN + Bi-directional LSTM), Fig. 8 (with the CNN + Uni-directional LSTM + Attention), and Fig. 9 (with the CNN + Bi-directional LSTM + Attention). Confusion matrix for different models indicates the best performances

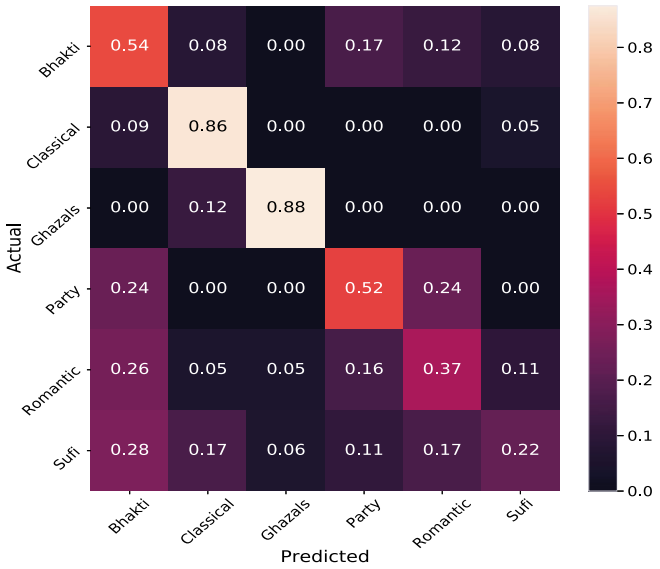

Fig. 5 Confusion matrix of model (CNN)

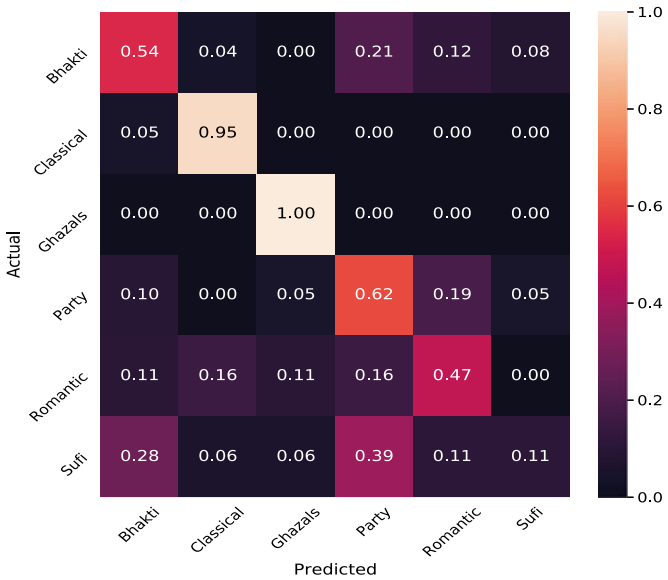

Fig. 6 Confusion matrix of model (CNN + Uni-LSTM)

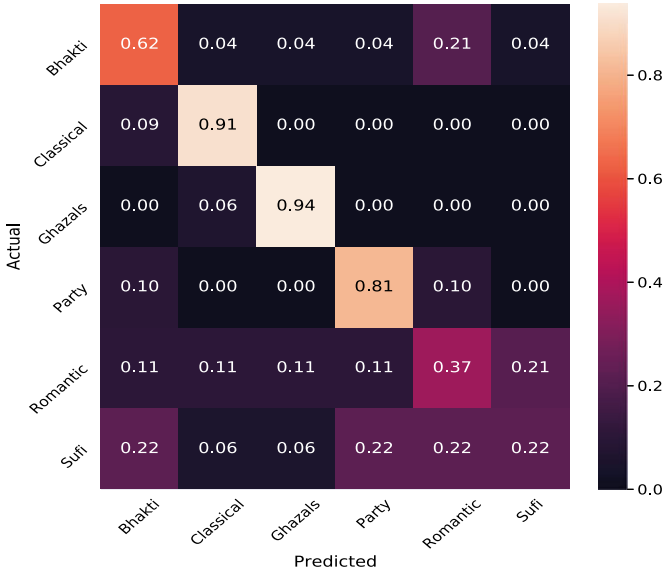

Fig. 7 Confusion matrix of model (CNN + Bi-LSTM)

for the music genre class, "*Classical*" and "*Ghazals*", while the worst performance is observed for "*Sufi*", such that a lot of them are classified under the "*Bhakti*" and the "*Romantic*" class.

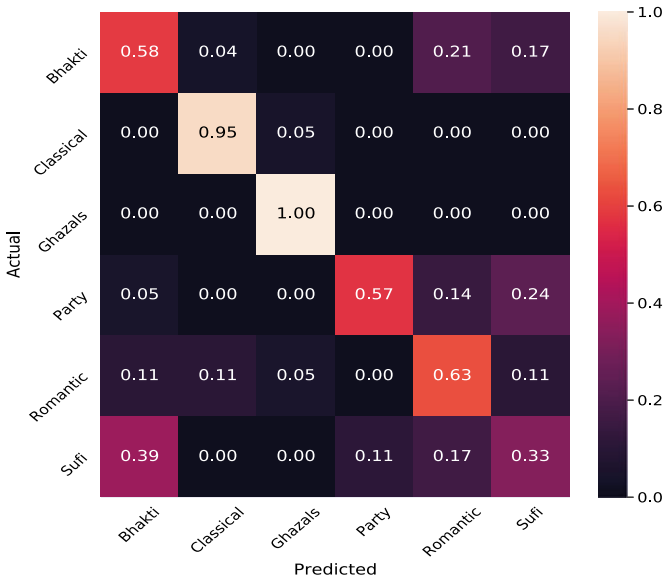

Fig. 8 Confusion matrix of model (CNN + Uni-LSTM + Attention)

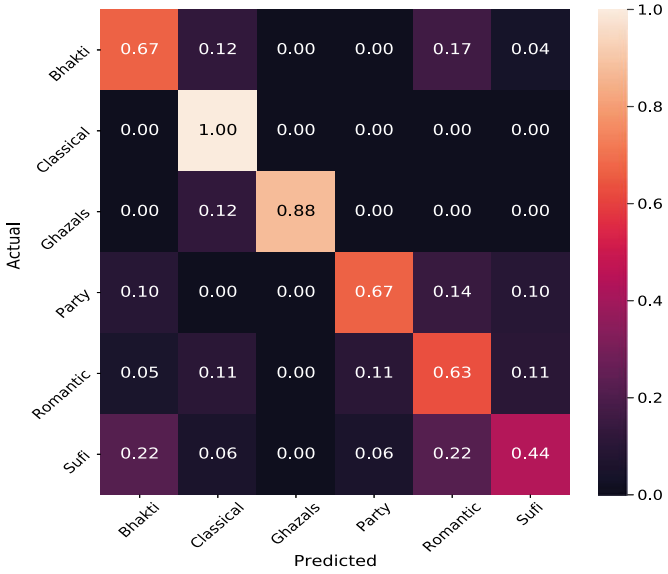

Fig. 9 Confusion matrix of proposed model (CNN + Bi-LSTM + Attention)

4 Conclusion

This work is for automatic genre identification of Hindi songs. There is no publicly available dataset. Therefore, a dataset is created for genre classification for Hindi songs. The deep learning architecture is developed to model the songs. Different features (Chroma, Pitch, Mel-spectrogram, and MFCC) are evaluated for the proposed architecture. Among all the features, MFCC produces the best F1-score of 71%.

References

1. Budhrani A, Patel A, Ribadiya S (2020) Music2vec: music genre classification and recommendation system. In: 2020 4th international conference on electronics, communication and aerospace technology (ICECA). IEEE, pp 1406–1411
2. Liang Dawen, Haijie Gu, O'Connor Brendan (2011) Music genre classification with the million song dataset. Machine Learning Department, CMU
3. Andrew Minkyu Sang (2020) Predicting musical genres using deep learning and ensembling. University of California, Los Angeles
4. Meng A, Ahrendt P, Larsen J, Kai Hansen L (2007) Temporal feature integration for music genre classification. IEEE Trans Audio Speech Lang Process 15(5):1654–1664
5. Bahuleyan H (2018) Music genre classification using machine learning techniques. arXiv:1804.01149
6. Chaudhary D, Singh NP, Singh S (2019) Genre based classification of hindi music. In: Innovations in bio-inspired computing and applications, Cham. Springer International Publishing, pp 73–82
7. Fu Z, Lu G, Ting KM, Zhang D (2010) A survey of audio-based music classification and annotation. IEEE Trans Multimed 13(2):303–319
8. Zhu Y, Kankanhalli MS (2006) Precise pitch profile feature extraction from musical audio for key detection. IEEE Trans Multimed 8(3):575–584
9. Bartsch MA, Wakefield GH (2005) Audio thumbnailing of popular music using chroma-based representations. IEEE Trans Multimed 7(1):96–104
10. Albawi S, Mohammed TA, Al-Zawi S (2017) Understanding of a convolutional neural network. In: 2017 international conference on engineering and technology (ICET). IEEE, pp 1–6
11. Khunarsa P, Lursinsap C, Raicharoen T (2010) Impulsive environment sound detection by neural classification of spectrogram and mel-frequency coefficient images. In: Advances in neural network research and applications. Springer, pp 337–346
12. Graves A, Fernández S, Schmidhuber J (2005) Bidirectional lstm networks for improved phoneme classification and recognition. In: International conference on artificial neural networks. Springer, pp 799–804
13. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv:1409.0473
14. McFee B, Raffel C, Liang D, Ellis DPW, McVicar M, Battenberg E, Nieto O (2015) librosa: audio and music signal analysis in python. In: Proceedings of the 14th python in science conference, vol 8, pp 18–25. Citeseer
15. Fahad MS, Deepak A, Pradhan G, Yadav J (2021) DNN-HMM-based speaker-adaptive emotion recognition using MFCC and epoch-based features. Circ Syst Signal Process 40(1):466–489
16. Ranjan A, Fahad MS, Fernández-Baca D, Deepak A, Tripathi S (2020) Deep robust framework for protein function prediction using variable-length protein sequences. IEEE/ACM Trans Comput Biol Bioinf 17(5):1648–1659

17. Chen Mingyi, He Xuanji, Yang Jing, Zhang Han (2018) 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Process Lett 25(10):1440–1444