INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Comparative Study of Mfcc and Mel Spectrogram for Raga Classification Using CNN

Dipti Joshi[1]*, **Jyoti Pareek**[2], **Pushkar Ambatkar**[3]

1 Research Scholar, Department of Computer Science, Gujarat University, India
2 Professor and Head, Department of Computer Science, Gujarat University, India
3 Student, Department of Computer Science, Gujarat University, India

OPEN ACCESS

Received: 06-09-2022

Accepted: 01-12-2022

Published: 20-03-2023

Citation: Joshi D, Pareek J, Ambatkar P (2023) Comparative Study of Mfcc and Mel Spectrogram for Raga Classification Using CNN. Indian Journal of Science and Technology 16(11): 816-822. https://doi.org/10.17485/IJST/v16i11.1809

* **Corresponding author**.

joshidipti1408@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To perform a comparative study of the results of feature extraction done using two different methods, Mfcc and Mel spectrogram, and determine which method is more effective for implementing the CNN algorithm. **Methods:** This study uses the CNN model to classify ragas according to Indian classical music. Feature extraction, which is a major operation in the Music Information Retrieval (MIR) process, is done using Mfcc and Mel spectrogram methods. The major ragas chosen as subjects for feature extraction are Yaman, Bhairav, Bhairavi, Multani, and Dhanashree. **Findings:** After comparison and examination of results achieved from both techniques, we could conclude that the CNN model using the Mel spectrogram method outperforms the CNN model using Mfcc. **Novelty:** The majority of the research, we discovered was on Carnatic music. In contrast to earlier research, this research takes a novel approach by conducting experiments on a variety of Hindustani classical ragas which are different from other studies. Researchers interested in music as well as application users would benefit from this study. Our proposed feature extraction approach will be useful for initializing the CNN algorithm, which will help aspiring musicians recognize ragas and classify songs based on these ragas.

Keywords: Raga identification; Music Information Retrieval; Feature extraction; Mfcc; Melspectrograms; CNN

1 Introduction

Indian classical music is one of the indissoluble aspects of Indian culture and society. A raga is the fundamental structure within Indian classical music, and it is a musical entity with its own musical personality. A raga is a blend of several distinctive swaras and can be identified using different methods like scale matching, Aaroh-Avroh pattern, Pakad matching, pitch class distribution, and Swara intonation.

Carnatic (South Indian) music and Hindustani (North Indian) music are the two types of Indian classical music (in Hindi, ”Bhartiya Shaastriya sangeet”) [1]. Raga is a concept shared by both music and dance. A raga is made up of seven swaras, one of which is ”SA RA GA MA PA DHA NI.” Swaras are melodies that are tuned at different frequencies, and the Raga is the basic grammar for Indian classical music composition and improvisation.

A composition of Hindustani classical music is known as a bandish, which literally means ”binding”. Each bandish consists of a unique blend of the central elements of Indian classical music. Swara or note, Aaroh-Avroh, Vadi-Samvadi, Gamakas, Pakad, Tala, and Thaat are all central elements that are used to identify a raga[2].

In classical music, raga identification is a challenging task for any researcher[3]. They have to face some issues such as:

- Understanding music is a challenging task that needs a high level of expertise.
- Many different instruments might be used during the composition of music.
- The raga notes are not in any particular order.
- Because of the many file formats, gathering musical data is challenging.

In the past, Carnatic raga classification has received a lot of attention, but Hindustani raga classification still needs a lot more effort. Due to the fast expansion of the digital music industry, the idea of automated music classification and identification has grown significantly in recent years[4]. In this study, various ragas from various Thaats have been used. The purpose is to compare the outcomes of feature extraction performed with the help of two alternative methods, the Mfcc and the Mel spectrogram, and ascertain which approach is more efficient for developing the CNN algorithm.

The remaining part of the paper is organised as follows: The methodology is discussed in Section 2. Section 3 is on results and discussion. The scope and future work are included in Section 4.

2 Methodology

2.1 Dataset

In our experiment, we have collected all the vocal and instrument audio files from the open-source platform YouTube. The dataset consists of songs sung by various musicians, a wide range of compositions, and a diverse variety of ragas. Firstly, we converted each audio clip to .wav format and then divided each into 15 second audio clips. We have approximately 2940 audio clips of classical ragas, including Yaman, Bhairav, Bhairavi, Multani, and Dhanashree. There are 584 audio clips of Yaman, 574 audio clips of Bhairav, 579 audio clips of Bhairavi, 603 audio clips of Dhanashree, and 600 audio clips of Multani. In the aforesaid experiment, we implemented both Mfcc and Mel spectrogram feature extraction approaches with CNN. The aim of this experiment was to compare, implement, and perceive the best approach.

2.2 Data Preprocessing

The effectiveness of the entire system can be significantly impacted by data pre-processing. The major goal of the pre-processing procedures is to effectively represent the audio input so that the deep learning models can extract the features quickly[5]. The audio files were sampled at a normal 44100 kHz rate. A mono audio signal is created by converting a stereo audio signal. The down sample rate for this signal was 11025 Hz.

2.3 Feature Extraction

Feature extraction means extracting meaningful characteristics from an audio signal before training any model. It's about how audio signals are processed or manipulated[6]. By translating digital and analogue signals, it eliminates undesirable noise and balances time-frequency ranges. It concentrates on sound-altering computational techniques. This paper introduces two important feature extraction techniques, namely Mfcc (Mel Frequency Cepstral Coefficients) and Mel-spectrogram, in detail.

2.3.1 MFCC

This is one of the most important approaches for extracting characteristics from an audio signal, and it is used frequently when working with audio signals. It simulates the features of human voice[7]. A signal's Mel-frequency Cepstral Coefficients (MFCCs) are a small group of characteristics (often 10–20) that concisely define the basic structure of a spectral envelope.

The Mel-frequency cepstrum (Mfc) is a short-term power spectrum representation of a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency in audio processing. The operations of Mel Frequency Cepstral Coefficients (Mfcc) include windowing the signal, determining the discrete Fourier transform (DFT) coefficients for each window signal, taking the log of the magnitude of the DFT, wrapping frequencies with the Mel scale filter, and then extracting the MFCC coefficients. The following diagram shows the process of Mfcc feature extraction (Figure 1).

A matrix called MFCC is extracted from each raga wave file. The feature set consists of the power spectrogram, mean of the MFCC, spectral centroid, zero-crossing rate, roll-off frequency, and spectral bandwidth. When we implemented Mfcc on the audio dataset, we received the Mfcc result as given in Figure 2.

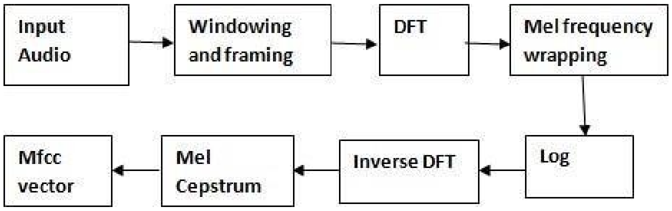

Fig 1. Mfcc feature extraction process

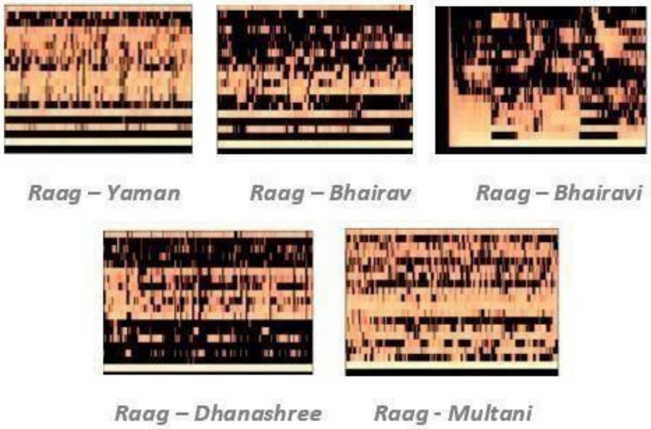

Fig 2. Sample Mfcc of each Raag

2.3.2 Mel-Spectrogram

The frequencies transferred to the Mel scale are known as the Mel spectrogram. Simply put, the Mel scale is the melody of any song. Mathematically, the Mel scale is the result of a nonlinear transformation of the frequency scale. This Mel scale is configured to be ”audible” to humans because the evenly spaced sounds on the Mel scale are equidistant from each other. It was discovered that converting audio files to images is a more effective audio pre-processing method than extracting numerical features[8]. The Mel Spectrogram is the result of the following process (Figure 3):

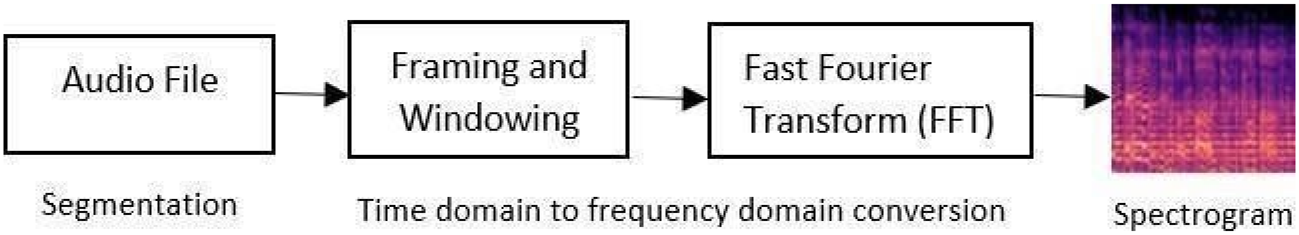
Audio File
Framing and Windowing
Fast Fourier Transform (FFT)
Segmentation
Time domain to frequency domain conversion
Spectrogram

Fig 3. Mel spectrogram extraction process

- To sample the next window, sample the input using windows with a sampling rate of 22050, a size of nfft = 2048, and a hop length of 512 each time.
- Compute the FFT (Fast Fourier Transform) for each window, convert the time domain to the frequency domain.
- Construct a Mel scale by separating the full frequency spectrum into n_mels = 128 equally spaced frequencies. in which distance is perceived by the human ear.
- Create a spectrogram for every window, decomposing the magnitude of the signal into its components, which corresponds to the frequencies on the Mel scale.

A time-frequency representation of a sound is produced by the Mel spectrogram, simulating the biological auditory systems of humans[9]. Figure 4 shows the spectrogram images that were created using the Librosa Python library.

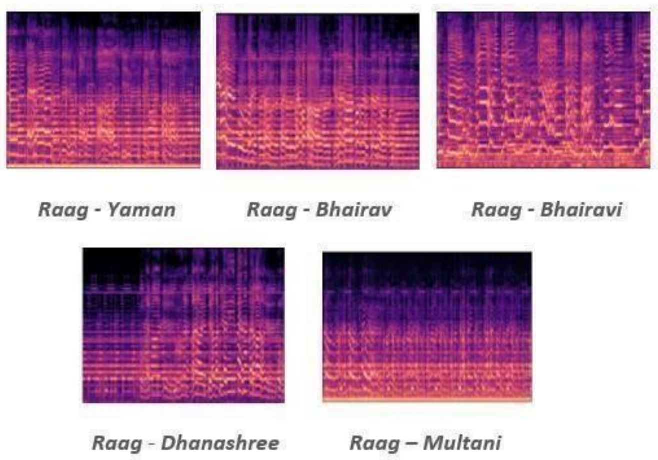
Raag - Yaman
Raag - Bhairav
Raag - Bhairavi
Raag - Dhanashree
Raag – Multani

Fig 4. Sample spectrogram images of each raag

2.4 Model

A CNN (Convolutional Neural Network) is a subset of a neural network that can extract a raga's specific characteristics from the prominent pitch values of a song[10]. From the raw pitch data, CNN can recognize the properties. In earlier experiments, the convolutional neural network (CNN) was first used to identify numeric data. With its evolution, CNN is applied for the implementation of image recognition and classification as well[11]. In the current research scenario related to audio datasets, diverse feature extraction techniques, including Mfcc and Mel spectrogram, are highly in demand. Due to the advancement of image classification techniques, an ascertained wide scope for the classification of music has emerged. The purpose of CNN is to map data from images and send it to an output variable for additional processing[12]. Here, specialized Python embedded packages are applied in our experiments, including Pydub for audio segmentation, Librosa for music analysis and information retrieval, and TensorFlow-Keras for interacting with neural networks.

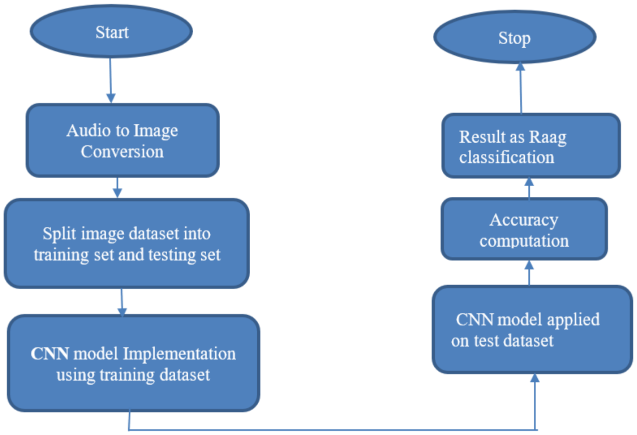
Start
Audio to Image Conversion
Split image dataset into training set and testing set
CNN model Implementation using training dataset
CNN model applied on test dataset
Accuracy computation
Result as Raag classification
Stop

Fig 5. Processflow of CNN model

Figure 5 depicts the process flow of the CNN model. Initially, the audio dataset was collected from open data sources. With the help of the Librosa library, the conversion of audio to a spectrogram image has been done. The entire dataset is divided into a training set and a testing set. On the training dataset, the CNN model is implemented. Once a model has been fully trained, the test dataset is used to evaluate the model. During the testing phase, the accuracy of the model can be analyzed, and if it is better, then it classifies the raga significantly.

We hereby propose a CNN-based approach to extracting features from an image of music audio. The model is designed using convolutional and max pooling layers, which are connected by the softmax layer explained in the Figure 6. Here is a brief

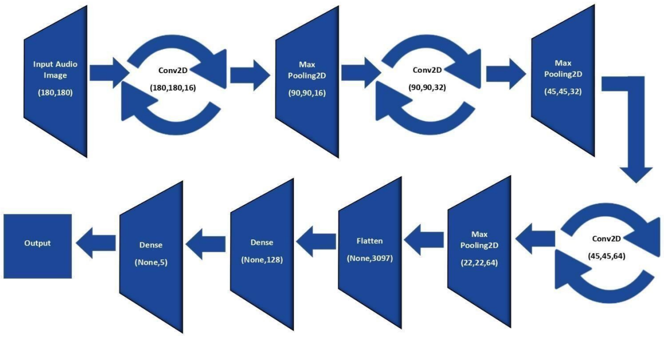

Fig 6. CNN Architecture

explanation of the CNN model components: In the input layer, an audio image made up of 180x180 pixels with RGB color code is fed to the next layer for processing. A convolutional layer is made up of a collection of learnable filters, or kernels. Three convolution layers are designed with some neuron combinations. This layer scans the pixels in the input image before creating the feature map. The convolutional layers have a stride of 1. Three max pooling layers with various neuron combinations are used. In each hidden layer, a model uses the "ReLU" activation function to activate neurons with positive values in prediction, whereas the final layer uses Softmax. Softmax normalizes the previous layer's output and computes the probability range between 0 and 1. Flattening data into a one-dimensional matrix is a useful technique. We employ two dense layers to classify images depending on the output of prior convolutional layers. In this experiment, we use the Adam (Adaptive Momentum) Multiclass Optimizer.

3 Result and Discussion

This section presents, comparison of various studies. Table 1, shows the work done by different authors.

Table 1. Comparison of various studies

Sr. No.	Authors	Algorithm	Features Extraction Method	Dataset	Accuracy
1.	Anand A.[1]	CNN	Pitch values Method	Carnatic	96%
2.	Joshi D, Pareek J, Ambatkar P.[2]	KNN SVM	Mfcc	Hindustani	98% 95%
3.	Vishnupriya S., Meenakshi K.[3]	CNN	Mfcc Mel spectrogram	Music Genre Dataset	47% 76%
4.	John S, Sinith M, RS S, PP L.[4]	CNN	Pitch detection algo.	Carnatic	94%
5.	Shah D, Jagtap N, Talekar P, Gawande K.[5]	CNN	Spectrogram	Hindustani	98.98%
6.	Bidkar A, Deshpande R, Dandawate Y.[6]	Ensemble bagged tree Ensemble KNN	Mfcc	Hindustani	96.32% 95.83%
7.	Patil N., Nemade M.[7]	KNN Linear Kernel SVM Poly Kernel SVM	Mfcc	GTZAN Dataset	64% 60% 78%
8.	Hebbar D., Jagtap V.[8]	1-D CNN 2-D CNN LSTM ANN	Mfcc Mel spectrogram	Carnatic (Pair of Ra gas)	97.4% 98.1% 97.54% 97%
9	Ghosal D., Kolekar M.[9]	CNN- LSTM	Mel Spectrogram	GTZAN Dataset	94.2%
10.	Dalmazzo D, Ramirez R.[10]	1-D CNN 2-D CNN CNN-LSTM	Mel-spectrogram	professional violinists Dataset	95.16% 84.30% 97.47%
11.	Rajan R, Sreejith S.[11]	CNN	Mel-spectrogram	Carnatic YouTube Dataset	F1 measure of 0.61

Continued on next page

Table 1 continued

12.	Phulmante V., Bidkar A., Mundada Y., Kulkarni P.[12]	CNN	Spectrogram Chroma STFT	Mfcc	GTZAN dataset	91% 72% 57%

From the table, it is clearly visible that authors have used different datasets but our dataset is absolutely different.When implemented on a dataset, CNN for both Mfcc and Mel spectrogram feature extraction approaches received the confusion matrix as shown in Figures 7 and 8. A confusion matrix depicts how well the classifier performed in the experiment when comparing (horizontal axis) expected outcomes and (vertical axis) actual outcomes. Here in the figure, five predicted and actual classes are included, such as class 0: Yaman, class 1: Bhairav, class 2: Bhairavi, class 3: Multani, and class 4: Dhanashree. In addition, we have depicted our result analysis in Figure 9 using a graph.

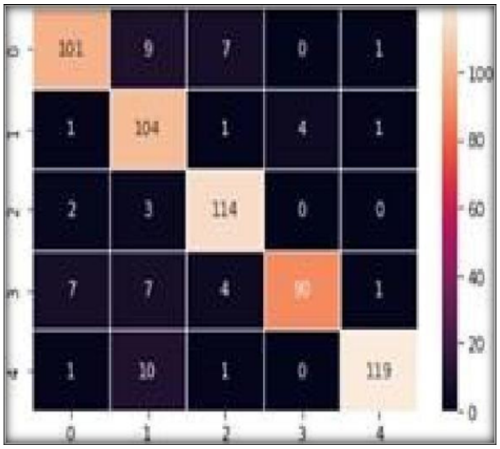

Fig 7. Confusion matrix of Mfcc

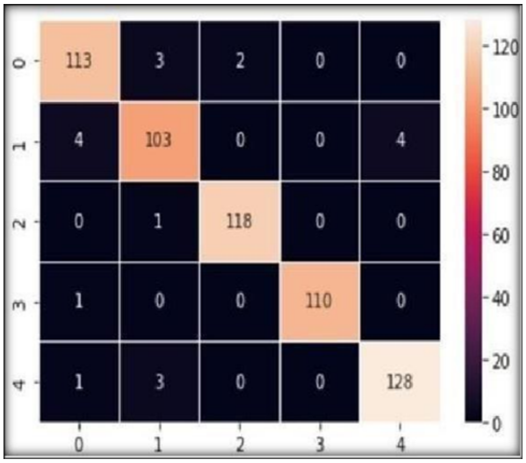

Fig 8. Confusion matrix of Mel spectrogram

It shows the comparison between CNN with Mfcc and CNN with Mel spectrogram. In terms of training accuracy, validation accuracy, and testing accuracy, the CNN achieves its overall performance for all 5 ragas, as illustrated diagrammatically in Figure 9. It has been shown that the accuracy of CNN using the Mfcc approach is 85% during testing, 89.79% during validation, and 97.62% during training. Whereas the accuracy for CNN using the Mel spectrogram approach during training is 98.81%, the validation accuracy is 96.78%, and the testing accuracy is 89%.

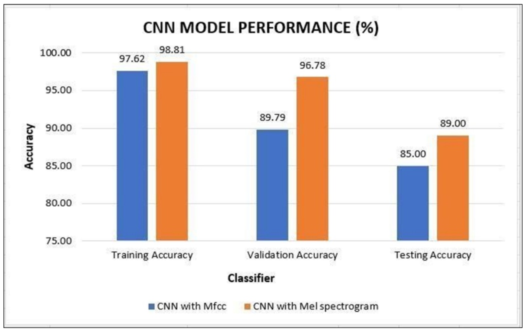

Fig 9. Model accuracy using Mel spectrogram and Mfcc

4 Conclusion

For Indian music information retrieval systems, raga identification is a crucial stage. The process of identifying a raga involves identifying distinctive notes and placing them in a predetermined order. In this paper, we have demonstrated an automatic technique for classifying and identifying selected ragas. The goal of this study is to compare two state-of-the-art approaches, CNN with Mfcc and CNN with spectrogram, simultaneously. The conclusion drawn is that developing a model as described can automatically classify and identify Ragas. It performs noticeably better with the Mel spectrogram. We have a distinctive and advanced research approach. As a result of the variety of raga datasets we selected, including Yaman, Bhairav, Bhairavi, Multani, and Dhanashree. We applied a strategy as we extracted features using both the Mfcc and Mel-spectrogram methods. Also, by using CNN, we conducted a comparison of the Mfcc and Mel-spectrogram.

It is also construed that the study entails and carries the potential to further investigate different ragas and increase the available dataset, aiming to achieve even higher performance in the future, as detailed exploration might also be conjoined with different algorithms. In future, researchers can also utilize the raga signal and one-dimensional CNN to perform raga classification.

References

1) Anand A. Raga Identification Using Convolutional Neural Network. *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*. 2019. Available from: https://doi.org/10.1109/ICACCP.2019.8882942.
2) Joshi D, Pareek J, Ambatkar P. Indian Classical Raga Identification using Machine Learning. 2021. Available from: https://ceur-ws.org/Vol-2786/Paper34.pdf.
3) Vishnupriya S, Meenakshi K. Automatic Music Genre Classification using Convolution Neural Network. *2018 International Conference on Computer Communication and Informatics (ICCCI)*. 2018. Available from: https://doi.org/10.1109/ICCCI.2018.8441340.
4) S J, M S, RS S, PP L. Classification of Indian Classical Carnatic Music Based on Raga Using Deep Learning. 2020. Available from: https://doi.org/10.1109/RAICS51191.2020.9332482.
5) Shah DP, Jagtap NM, Talekar PT, Gawande K. Raga Recognition in Indian Classical Music Using Deep Learning. *Artificial Intelligence in Music, Sound, Art and Design*. 2021;p. 248–263. Available from: https://doi.org/10.1007/978-3-030-72914-1_17.
6) Bidkar A, Deshpande R, Dandawate Y. A North Indian Raga Recognition Using Ensemble Classifier. *International Journal of Electrical Engineering and Technology (IJEET)*. 2021;12(6):251–258. Available from: https://sdbindex.com/Entry/both/18635.
7) Patil N, Nemade M. Music Genre Classification using Mfcc, K-NN and SVM classifier. *International Journal of Computer Engineering in Research Trends*. 2017;(4):43–47. Available from: https://ijcert.org/ems/ijcert_papers/V4I206.pdf.
8) Hebbar D, Jagtap V. A Comparison of Audio Preprocessing Techniques and Deep Learning Algorithms for Raga Recognition. 2022. Available from: https://doi.org/10.48550/arXiv.2212.05335.
9) Ghosal D, Kolekar MH. Music Genre Recognition Using Deep Neural Networks and Transfer Learning. *Interspeech 2018*. 2018. Available from: https://doi.org/10.21437/Interspeech.2018-2045.
10) Dalmazzo D, Ramirez R. Mel-spectrogram Analysis to Identify Patterns in Musical Gestures: A Deep Learning Approach. 2020. Available from: https://www.researchgate.net/publication/345671783_Mel-spectrogram_Analysis_to_Identify_Patterns_in_Musical_Gestures_a_Deep_Learning_Approach.
11) Rajan R, Sreejith S. Raga Recognition in Indian Carnatic Music Using Transfer Learning. 2021. Available from: https://doi.org/10.37394/232019.2022.9.2.
12) Phulmante V, Bidkar A, Mundada Y, Kulkarni P. Recognition of music genres using deep learning. *International Research Journal of Engineering and Technology 2022*;p. 5–5. Available from: https://www.irjet.net/archives/V9/i5/IRJET-V9I5253.pdf.