Deep Learning-Based Classification of Indian Classical Music Based on Raga

Conference Paper · March 2023

DOI: 10.1109/ISCON57294.2023.10111985

4 authors:

Anupam Singha
R. P. Saha University

Rajalakshmi N R
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology

Arun Pandian J
Vellore Institute of Technology University

Saravanan Srinivasan
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology

2023 6th International Conference on Information Systems and Computer Networks (ISCON) | 979-8-3503-4696-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISCON57294.2023.10111985

Deep Learning-Based Classification of Indian Classical Music Based on Raga

1st Anupam Singha
Department of Computer Science and Engineering
Vel Tech Ranagrajan Dr Sagunthala
R&D Institute of Science and Technology
Avadi, Chennai, 600062, India
anupameos@gmail.com

2nd Dr. N R Rajalakshmi
Department of Computer Science and Engineering
Vel Tech Ranagrajan Dr Sagunthala
R&D Institute of Science and Technology
Avadi, Chennai, 600062, India
drnrrajalakshmi@veltech.edu.in

3rd Dr. Arun Pandian J
Department of Computer Science and Engineering
Vel Tech Ranagrajan Dr Sagunthala
R&D Institute of Science and Technology
Avadi, Chennai, 600062, India
jarunpandian@veltech.edu.in

4th Dr. S. Saravanan
Department of Computer Science and Engineering
Vel Tech Ranagrajan Dr Sagunthala
R&D Institute of Science and Technology
Avadi, Chennai, 600062, India
drsaravanans@veltech.edu.in

***Abstract*—Classical music is an integral part of Indian culture's lineage. The raga serves as the foundation for all Indian classical music, including both Hindustani and Carnatic music. The term "raga" is typically used to refer to the melodic structure in Indian classical music. Traditional approaches to the classification of ragas are time consuming and inefficient. In this work, a convolutional neural network was proposed for raga classification. The proposed convolutional neural network was trained on 70 classes of ragas. The proposed convolutional neural network takes the spectrograms of the audio note and identifies the raga of the note. The proposed convolutional neural network model achieved a precision of 97.9% on raga classification. The performance of the proposed convolutional neural network was compared with the traditional classification techniques using standard performance metrics such as precision, recall, F1 score, and AUC-ROC. The comparison results show that the classification performance of the proposed convolutional neural network was superior to the state-of-the-art machine learning techniques.**

Keywords —Music Genre classification, Musical features, Audio signal processing, Music information retrieval, Raga identification, Deep learning

I. Introduction

Music is widely perceived as the medium of expression. It is an artistic medium that delivers an expressive and widely accepted message that speaks to all individuals, irrespective of race or origin [1]. Every genre of music has its own note, harmony, tempo, melodies, and rhythm. Carnatic and Hindustani music these are two major styles of Indian Classical Music, which supports the vast of Indian subcontinental music. Raga provides a melodic basis for both composition and experimentation, an integral part of Indian classical music. The Northern Classical Music tradition is known as Hindustani, whereas the Southern Classical Music tradition is known as Carnatic. The most significant components of Indian classical music are raga and talam. In a composition, Talam specifies where the syllables are put and the rhythm of the music. Raga structures distinguish Hindustani Classical Music and Carnatic Classical Music. According to Klapuri, a vocalist's musical note is considered to be a distinguishable frequency component (usually known as pitch) [2]. An interval is calculated by taking the underlying frequencies of two notes by their ratio. The seven musical notes Sa, Ri, Ga, Ma, Pa, Da, and Ni have frequencies and that also can be broken into semitones or microtones [3]. There are three different scale types used in Hindustani and Carnatic music: a 12-note scale used in Hindustani music; a 16-note scale used in Carnatic music, and a 22-note microtone scale [2]–[5]. Carnatic music has been shown to have 12 different frequency components. Despite the fact that these two genres of music share a similar structure, they have grown in various ways due to diverse cultural influences [6]. A Raag is a musical note combination that is pleasant when sung or played on an instrument. By combining the twelve semitones in a particular pattern, raga may convey sentiment in a song. As a result, to identify a Raga the occurrence or missing of notes is essential. Although in many melodies that contain the same set of notes, arohana-avarohana patterns are important. Arohana-avarohana patterns are note progressions that are dropping and rising. It tells about the many note transformations that might occur in a raga [7]. There are around 100 fairly frequently used ragas, and classifying them is a challenging task. The majority of research in this field has focused on Western music, with only a few studies attempting to research Indian classical music [2], [7], [8].

The purpose of this work is to construct a deep learning based raga classification model for Indian classical music that can identify ragas in real-time. There have been several methods of identifying the raga from melodies. The majority of classical raga identification techniques rely on basic frequency detection methods to retrieve notes, which are subsequently mapped into matching ragas. The fundamental frequency of a sound determines its pitch, which is determined both in the time and frequency domain. The Hidden Markov Models (HMM), Pitch Monitoring, and the Basic Rule-Base Heuristic were among the early approaches. These traditional approaches are quite complex and computationally inefficient. The accuracy of classification and processing efficiency may be vastly enhanced using deep learning. The advancement of deep learning in different fields, especially in Speech Recognition (SR), Natural Language Processing (NLP), and Computer Vision (CV) [9], has inspired researchers to use deep learning-based approaches to classify music genres. Several techniques use unique features or groups of features from the deep learning model to classify music or sound. This paper proposed a deep learning-based convolutional neural network model to classify raga, named the 1D-CNN model, and a preprocessing pipeline to process the audio data for the proposed model.

II. Literature Review

Researchers looked into how retrieving melodies functioned in the context of Western music. There is very little work on Carnatic and Hindustani music. Only a few studies on Swara pattern recognition and the identification of singers have been conducted in Carnatic music [10], [11]. Simultaneously, significant research is undertaken in Hindustani music to identify the Ragas. An HMM based model created by the authors of [12], identified 2 different Ragas in Hindustani Classical Music. Hindustani and Carnatic music's raga patterns differ significantly [10]. Artificial neural networks have a wide range of uses, and Music Information Retrieval is no exception. The results of the research are assessed using several data sets gathered by the researchers.

A. Traditional Approaches

Chakraborty et al. [13] retrieved the svaras from the music and correlated them to the raga's already documented svaras. This approach is based on the svaras, which are quite fundamental features. Overall, 94% accuracy in identifying tonic and other notes obtained using the proposed technique. Though it may detect melodic characteristics of music to a certain level, it does not take into account the temporal aspects of the same. Furthermore, while not always the same raga, certain ragas contain the same grouping of svaras in Indian Classical Music.

In [14], to identify the Hindustani ragas they used the Hidden Markov Model (HMM). The author developed two heuristics: notation period heuristics and pitch contour heuristics. They use an n-gram model to identify repeating note sequences. The shortcomings of this model are the much less data collection and low note transcription accuracy. The model's overall accuracy is stated to be 77%.

[15] utilised arohana and avrohana, which are ascending and descending svara progressions. Authors addressed the temporal characteristics of the music using arohana and avrohana.To represent the music discretely, they used quantized pitch vectors. It's a pretty easy and quick method. Because of the discrete representation, the model is not exact.

In [10], the authors used a vector space model for describing audio signals in their study. To identify the ragas, they developed a prediction model identical to the vector space models from Natural Language Processing. As a result of using this model, raga-phrases can be identified. For numerous computational tasks, such phrases as melodic elements with semantic value can be utilised as a database. A byproduct of this approach is the identification of raga styles.

In [11], Time-Delayed Melody Surface (TDMS) generated improved results by performing preprocessing procedures such as Tonic Normalization, surface generation to produce TDMS, and Predominant Melody Estimation. TDMSs are easy to set up, efficient to compute, and also have a musical relevant interpretation. The model couldn't tell the difference between ragas that had a similar collection of svaras and phrases.

There are two stages proposed method by [16] for raga identification. Using the characteristics retrieved from the pitch histogram, the first phase, tonic, and raga were calculated individually. Phrase two adjusts the raga and tonic according to the information obtained from the notes. The characteristics obtained from the probability density function within the track on pitch values are utilised to detect raga in this study. The suggested approach is effective in identifying ragas created with various figures of notes and sets of notes. For each track, Melodic Frequency Cepstral Coefficients (MFCC) are computed at each end using speech signal processing methods. The input signal is a single melodic, vocal note. The suggested technique is validated using a large data set containing a range of music samples. Carnatic music has low accuracy due to its very sophisticated gamakas and improvisations.

The work [17], proposed a new method for identifying ragas. This involves retrieving the pitch vector and utilising an existing records to do a locality searching in the raga's region. The approach seems to be very flexible, with the ability to recognise up to 3000 ragas. Most crucially, Locality Sensitivity Hashing avoids the scaling concerns that plague previous techniques. The precision is compromised, and comparable ragas are difficult to distinguish.

In [18], proposed that similar ragas might be identified using note-embeddings. The notations were used to study these embeddings. The bandish notation of the compositions was used by the authors to train the recurrent neural network (RNN) as well as learn the note-embeddings. These cosine relationship embeddings of two matrices can be used to determine a similarity among two records.

B. Deep Learning Based Approaches

A deep learning architecture that efficiently trains and generalises on a small quantity of data is preferable. In general, architecture that generalises well and operates well on new inputs is difficult to handle and require technical work experience. The difficulty of building specialised architectures for a certain task is solved when researchers deploy enough deep architectures. These deep architectural models are applicable to a wide range of jobs. Using transfer learning, the model's performance on related tasks can improve over time given the learned weights from the architecture for which the original task was prepared. Existing raga identification works may be divided into two main categories [16]: (1) with explicit note information, and (2) without explicit note information.

In [19], the author presented an artificial neural network (ANN)-based method for raga identification in Carnatic music that uses note sequence information. Arohana-avarohana sequences are classed either linear or non-linear according on which pair of notes are included in the sequence. Twenty ragas are used to execute the process, and each one has 3 to 5 songs. In this study, raga identification accuracy was shown to be 95% accurate. A limited dataset was used to train the model.

The author of [12] built a model on the CompMusic dataset [10], [11] using the key pitch frequencies of a song with a Convolutional Neural Network (CNN) which is capable of recognising the special features of a raga. CNNs are used to analyse visual data of pitch vectors by feeding the model with grayscale visualizations. The accuracy of the research was 85.6% using the collection containing 11 distinct ragas and 96.7% with a certain collection containing 5 different ragas. When evaluated on associated ragas, the model's accuracy was about 50%.

The author [20] introduced PhonoNet, a multi-stage deep learning model. Due to vocal phrases throughout the songs, they converted the recordings in mel-chromagrams and used the Short Time Fourier Transform. First, CNN was trained to recognise ragas using smaller pieces of recordings from the dataset [10], [11], whereas LSTM recognises the rhythmic and melodic components within the songs and can recognise the raga. To recognise the raga's subtleties, CNN trained first to discover the proper filters. Only Hindustani Classical voice music was used to test this work.

In [21], audio source separation methods were used to preprocess the audio. With Mad Twinnet, the authors were able to isolate vocals from the mixed audio. Following tonic normalisation along with pitch monitoring, the pitch data was passed towards an LSTM, which has attention layers. On the CompMusic dataset [10], [11], the model "SRGM1" achieved an accuracy of 97% on the subset of 10 ragas. Furthermore, the researcher developed a model, "SRGM2," that deals with the sequence ranking issues. So when samples include almost no detail, the model underperforms, that related with occurrences like as lengthy breaks throughout the song, holding a certain note over such extended period of time, and so on.

The study [22] introduced a combination of CNN and LSTM-based approaches for identifying raga in audio through the use of raw spectrograms. Across a collection of 10 ragas from the CompMusic dataset [10], [11], the suggested model obtained 98.98 percent testing accuracy. They extensively examined how different hyperparameters affected outcomes. Compared to most of the earlier work, this model is more accurate in classifying the associated ragas.

III. Methodology

Using deep learning techniques and signal processing, difficult tasks such as raga classification from music can be accomplished. The audio samples may be shaped into sequences that deep learning models can understand with the right preprocessing. As a result, the raga recognition issue may be reduced to a sequence classification problem. This research presents a method for preprocessing audio inputs using the Data Preprocessing pipeline. Before being fed to a deep learning model, the raga is preprocessed as an audio input. The model's result will be the raga prediction for such audio input.

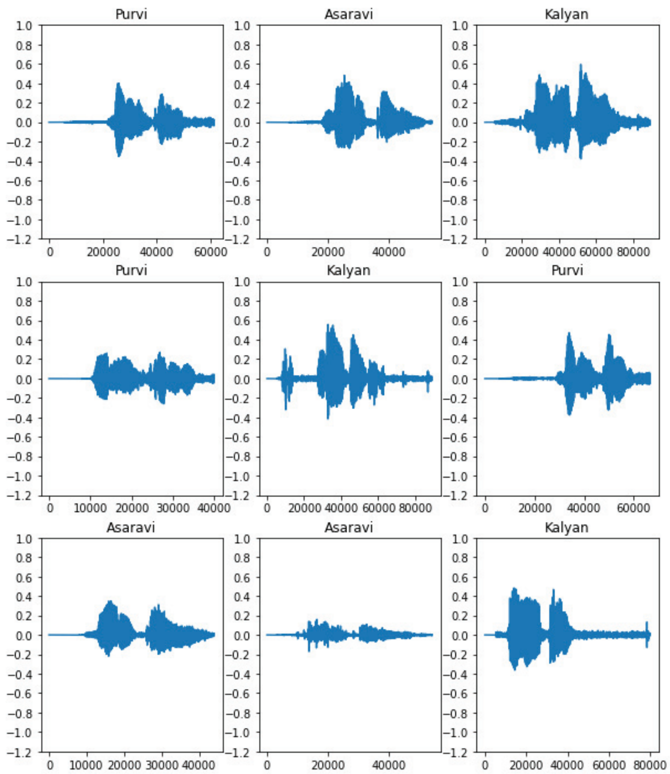

Fig. 1. Audio waveforms of various ragas used in this study

The Carnatic music dataset and the Hindustani music dataset are two large datasets that makes up the CompMusic dataset [10], [11] and were chosen for this work, each representing an unique musical tradition. The Carnatic dataset was compiled with Hindustani music in the Carnatic style. The test dataset comprises around 7470 minutes of recordings of 40 distinct ragas. There are 12 music samples in each Raga class. Hindustani Music Raga Dataset consisting of about 6972 minutes of audio files of 30 common ragas. Each raga in this collection has 10 recordings, however the length of those recordings may vary. The proposed model and preprocessing technique work well with the CompMusic dataset.

A. Data Preprocessing

The performance of the whole system is greatly driven by data preprocessing. The major goal of the preprocessing phases is to determine the optimal way to encode audio input, thus the features can be effectively extracted by the deep learning model. The audio file will be processed by the preprocessing pipeline, which will provide spectrogram visuals. In the preprocessing pipeline, the Down Sampler, Padding Silencing, Short Time Fourier Transform Unit, and Spectrogram Unit are employed.

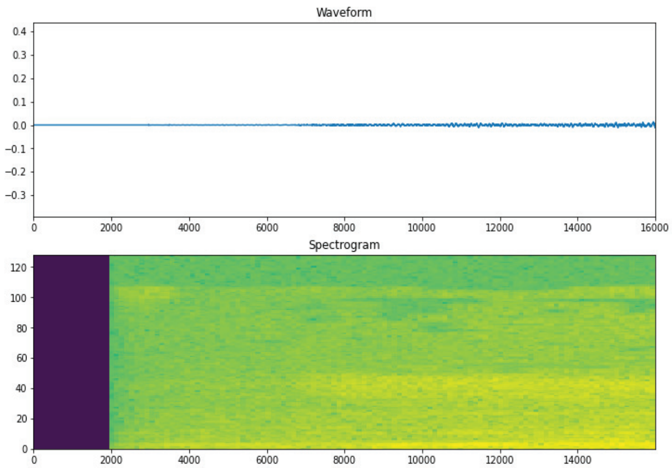

Fig. 2. Spectrogram image of the raga audio after down sampling and padding silence

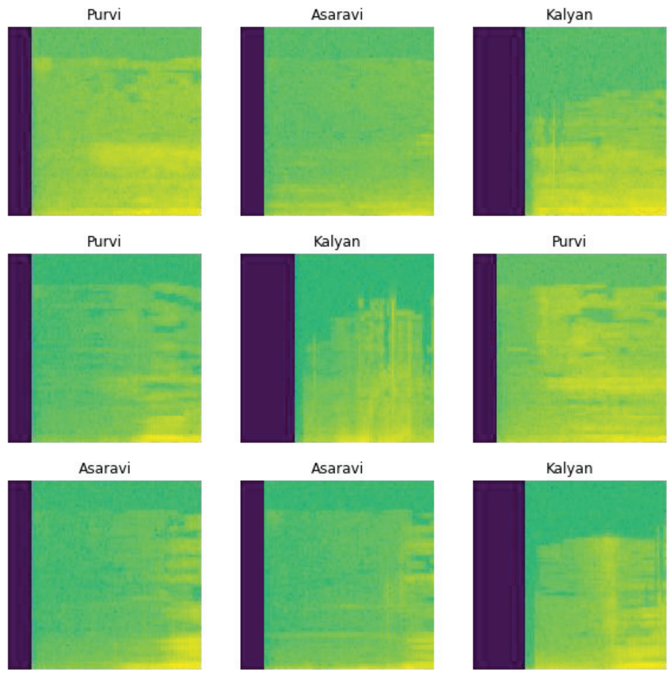

Fig. 3. Spectrogram images of various ragas

B. Model Architecture

From the preprocessing pipeline data, the model is fed, which provides a raga prediction. It is required to have a deep learning model that is capable of recognising the patterns of these features after extracting them through spectrograms. Convolutional Neural Networks (CNN) excel in capturing and representing features in a single vector. Unlike a standard neural network, a CNN's neurons are produced in such a 3-D structure constructed in a 3-D volume. Each layer's neuron is linked to a narrow fragment of the layer before it at a time. It thus reduces the number of characteristics they should take into consideration while processing images and enables them to cope with bigger data sets. In most cases, CNN is used to generate an encoder, which is then used to encode and retrieve image statistics. These characteristics can be learned by CNN from spectrogram images of the audio dataset. A sequence of feature vectors can be produced by utilising the CNN model to transform sequential audio spectrograms into feature vectors. The proposed CNN model is tuned and trained using audio spectrogram images for feature extraction and classification.

- **Convolutional Neural Network:** The excessively huge number of parameters is one of the key issues when working with image data. When working with image data to develop a deep learning model, the problem becomes even worse, since overfitting may occur. This work prefers max-pooling to keep the number of parameters and processing down as it provides the largest value in a specific window over the input layer. A kernel-based convolution layer, max-pooling, normalisation, and dropout layers are implemented. Figure 4 illustrates the layered structure of the single convolutional layer CNN model. The CNN model fed the input image in the shape of (Width–124, Height–129, Channels-1) in the form of a tensor. Before batch normalisation, the tensor's form was scaled to (32, 32, 1). Following that, a convolution

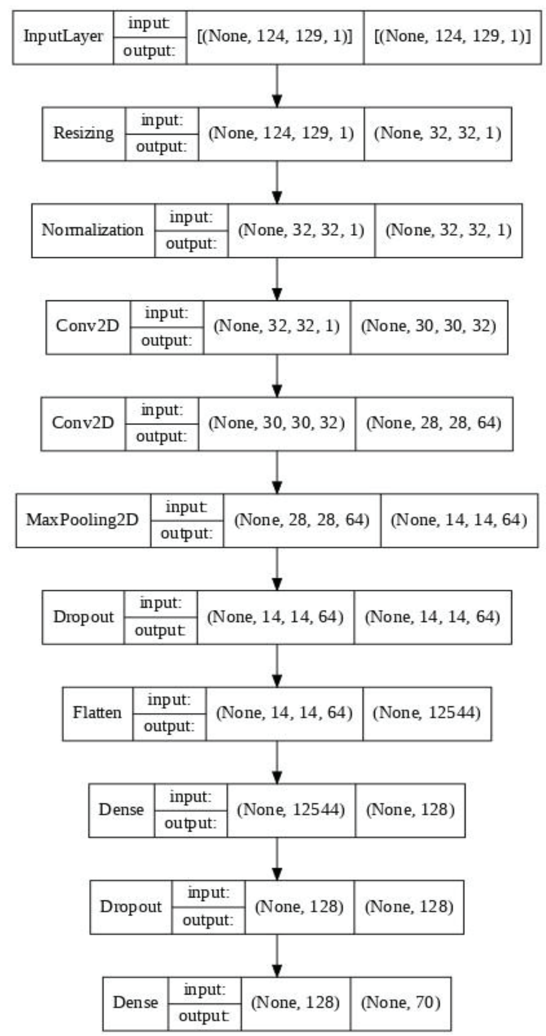

Fig. 4. The layered structure of the single convolutional layer CNN model

layer with a kernel size of 3 is used twice. The feature dimensions were compressed using max-pooling with a stride of 2 and dropout is applied. The input'sfeatures are flattened out of the tensor into a single vector with the shape 12544. It is termed the "feature vector". Following this, a completely linked hidden layer that is a design layer of 128 neurons is employed. Then a dropout applied again. Depending upon the quantity of ragas within the dataset, the last fully connected layer is next linked with an output layer of the same size. The 1D-CNN model is utilised for raga classification after it has been trained. Moreover, the model's final dense layer is utilized as a feature vector of the test audio recordings for raga categorization. The 1D-CNN model was trained using data from all 70 ragas. The test accuracy of the 1D-CNN model is 97.8%. The efficiency of the 1D-CNN model for the classification of raga is shown in Figure 5.

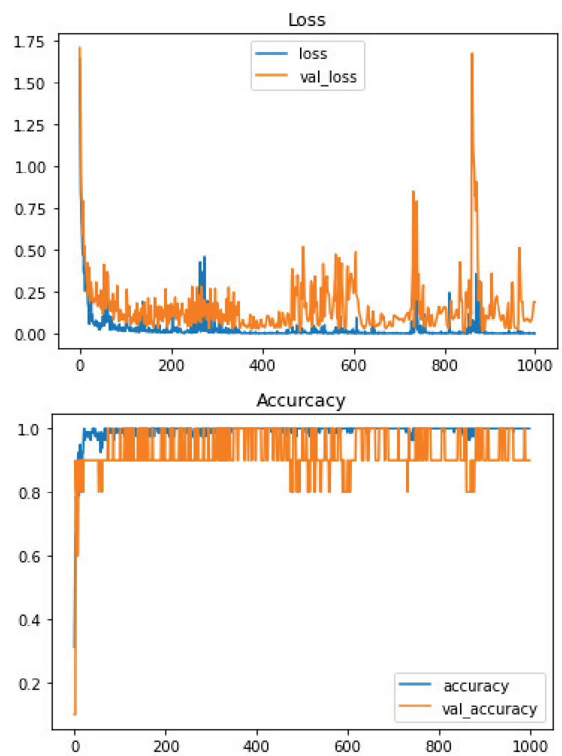

Fig. 5. The performance of the 1DCNN model

IV. Results And Discussions

To classify music [23] or audio [24] with deep learning, different approaches make use of various characteristics or groups of features. These techniques give more preference to the structure of the deep learning model than to the diversity of the features that are included and their robustness. Deep learning-based systems typically require large amounts of data and significant computational power to build a classification model. High-dimensional musical features cost more to run and occupy more storage space when fed to deep learning-based systems. Furthermore, it's possible that the provided musical feature failed to fully capture the implicit musical patterns for the task given. Enormous possibility remains for Indian Classical Music Information Retrieval systems using deep learning based techniques. The complexity of the raga recognition problem can be resolved using significant advances in signal processing and deep learning.

Based on the raga identification task's applied musical as well as non-musical features, the deep learning models' efficiency is observed. Traditional techniques that incorporate musical characteristics such as Chromagram and Tonnetz perform poorly in terms of classification accuracy. Overfitting is an issue with the models that use these features. Furthermore, low-level features such as nonmusical elements generally performed exceptionally well across deep learning models, with the exception of a few examples where low performance is seen. Considering the key performance metrics with the original input images, the suggested 1D-CNN model is tested with advanced classification algorithms.

The proportion of correctly reported values with all predicted values is referred to as precision. Figure 6 demonstrates the precision value of the 1D-CNN and other modern machine learning techniques. The precision of the 1D-CNN is significantly higher compared to other approaches.

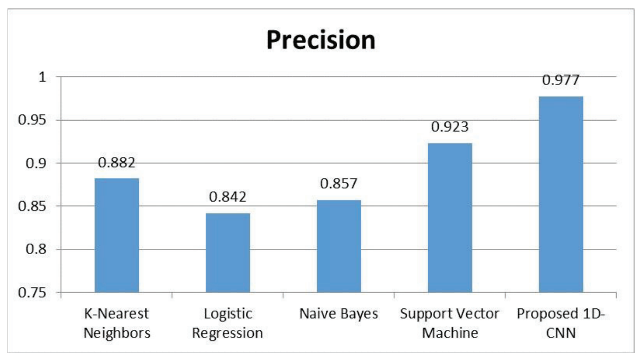

Fig. 6. Precision value comparisons with all modern techniques and the proposed 1D-CNN

The ratio of accurately predicted values to all known based on a given particular class is defined as recall. Figure 7 displays the recall value of the 1D-CNN, which is greater than that of the other techniques.

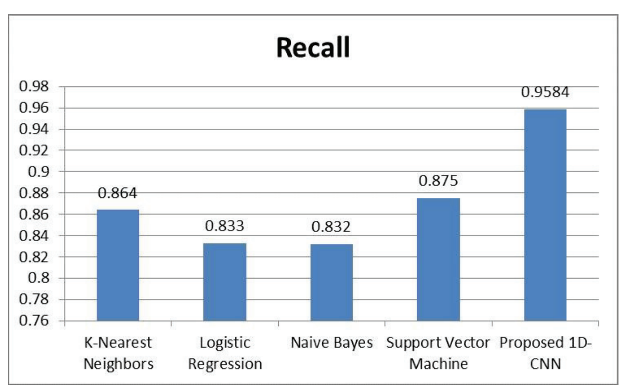

Fig. 7. Recall value comparisons with all modern techniques and the proposed 1D-CNN

The most essential performance statistic is the F1 score, which is determined as the weighted average of precision and recall. Figure 9 illustrates how the developed 1D-CNN model outperformed the other machine learning models with an F1-Score of 0.96.

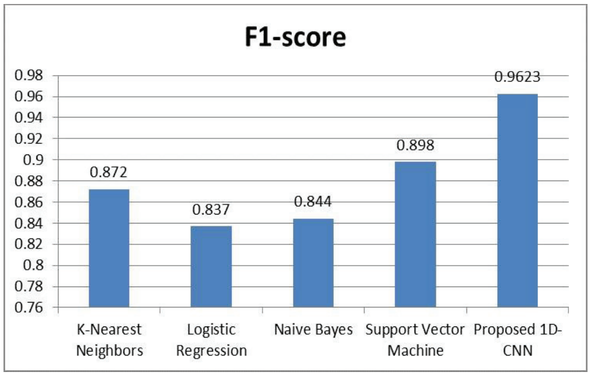

Fig. 8. F1 score comparisons with all modern techniques and the proposed 1D-CNN

A receiver operating characteristic (ROC) curves are frequently used to evaluate how well any model performs when challenged with the classification issue at various threshold levels. In essence, ROC is a probability curve. Keeping in mind the amount of data used for training it can be clearly seen that our proposed system produced comparable results with maximum AUC difference of 0.208.

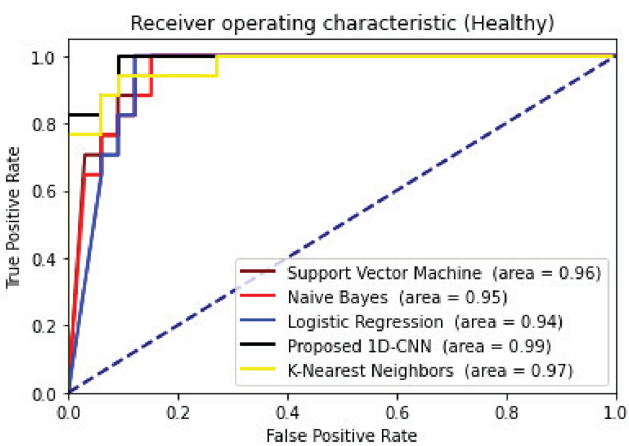

Fig. 9. ROC curve of all modern techniques and the proposed 1DCNN

The results of all performance metrics imply that the 1D-CNN model outperformed the other proposed stateof-the-art machine learning approaches. The model may be more accurate in predicting the raga at slower periods of the performance, but if the tempo of the performance increases, the model may become less accurate. However, the model will be able to recognise the raga. Because this research is focused on Indian classical music, improved data processing pipelines might significantly improve the model's performance. The model will be able to make better and more consistent predictions as a consequence of the sound source separation. The source separation may significantly increase the model's performance. The system could potentially use the frequency lines of instruments like the sitar, harmonium, or percussion instruments like tabla or mridangam to identify patterns. In terms of identifying unique ragas, the proposed model is significant to existing techniques.

V. Conclusion

In this work, an extensive literature review on this subject has been done. This work studied and compared different methods which can solve this problem. Based on the knowledge gained in this field, suitable deep learning frameworks and datasets are chosen for training our model. The proposed method is designed based on a large dataset comprised of a bunch of different samples.

Raga identification entails recognising distinct notes in a predefined sequence. It includes the analysis of both music and speech. This classification can provide a good picture of the field of Indian music. This classification approach may also be used to compare the raga's similarity in various forms. Music research is a new field that uses a variety of computer tools to explore different types of music. In Indian classical music, the portrayal of melody and its similar features are extremely important.

The music industry has changed dramatically during the previous decades. An enormous amount of music content provides a huge categorization and archiving difficulty. We have access to a massive amount of digital music content via internet streaming platforms like SoundCloud, Spotify, Youtube, Apple Music, Gaana etc. Computational musicology is a novel concept that studies approach to understanding different genres. It contributes to the study of music and the creation of a technical foundation for this kind of art. Music analysis helps in perceiving the culture and legacy of the civilization in which it arose.

The humdrum-based sequences might be used in future works. Applications for identifying raga from audio include, and aren't really limited to educational, content based filtering, and other fields. This really can be used for a variety of tasks, such as musical therapy and the synthesis of musical signals. Numerous downstream uses, such as storing big music libraries and generating music recommendations, might benefit from its use.

References

[1] A. Aljanaki, "Anna aljanaki automatic musical key detection master's thesis (30 ects)," 2011.

[2] A. Klapuri and M. Davy, "Signal processing methods for music transcription," 2006.

[3] Y. Jia, X. Chen, J. Yu, L.-m. Wang, Y. Xu, S. Liu, and Y. Wang, "Speaker recognition based on characteristic spectrograms and an improved self-organizing feature map neural network," *Complex Intelligent Systems*, vol. 7, 06 2020.

[4] S. Manjabhat, S. Koolagudi, K. Rao, and P. Ramteke, "Raga and tonic identification in carnatic music," *Journal of New Music Research*, vol. 46, pp. 1–17, 05 2017.

[5] B. Kumaraswamy and P. G. Poonacha, "Recognizing ragas of carnatic genre using advanced intelligence: a classification system for indian music," *Data Technol. Appl.*, vol. 54, pp. 383– 405, 2020.

[6] A. Vidwans, K. K. Ganguli, and P. Rao, "Classification of indian classical vocal styles from melodic contours," 2012.

[7] G. E. Poliner, D. P. W. Ellis, A. F. Ehmann, E. Gomez, S. Streich, and B. Ong, "Melody transcription from music audio: Approaches and evaluation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1247–1256, 2007.

[8] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, “A tutorial on onset detection in music signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[9] N. Rajalakshmi and K. Saravanan, “Traffic violation invigilation using transfer learning,” in *2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC)*, 2022, pp. 286–292.

[10] S. Gulati, J. Serrà, K. K. Ganguli, S. Sentürk, and X. Serra, “Time-delayed melody surfaces for r¯aga recognition,” in *ISMIR*, 2016.

[11] S. Gulati, J. Serrà, V. Ishwar, S. Sentürk, and X. Serra, “Phrasebased rAga recognition using vector space modeling,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 66–70.

[12] A. Anand, “Raga identification using convolutional neural network,” in *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, 2019, pp. 1–6.

[13] S. Chakraborty and D. De, “Object oriented classification and pattern recognition of indian classical ragas,” *2012 1st International Conference on Recent Advances in Information Technology (RAIT)*, pp. 505–510, 2012.

[14] G. Pandey, C. Mishra, and P. Ipe, “Tansen: A system for automatic raga identification,” in *IICAI*, 2003.

[15] V. Kumar, H. Pandya, and C. V. Jawahar, “Identifying ragas in indian music,” *2014 22nd International Conference on Pattern Recognition*, pp. 767–772, 2014.

[16] S. S. Manjabhat, S. G. Koolagudi, K. S. Rao, and P. B. Ramteke, “Raga and tonic identification in carnatic music,” *Journal of New Music Research*, vol. 46, no. 3, pp. 229–245, 2017. [Online]. Available: *https://doi.org/10.1080/09298215.2017.1330351*

[17] G. Padmasundari and H. A. Murthy, “Raga identification using locality sensitive hashing,” *2017 Twenty-third National Conference on Communications (NCC)*, pp. 1–6, 2017.

[18] J. C. Ross, A. Mishra, K. K. Ganguli, P. Bhattacharyya, and P. Rao, “Identifying raga similarity through embeddings learned from compositions’ notation,” in *ISMIR*, 2017.

[19] S. Shetty and K. K. Achary, “Achary. raga mining of indian music by extracting arohana-avarohana pattern,” *International Journal of Recent Trends in Engineering*, p. 2009.

[20] S. Chowdhuri, “Phononet: Multi-stage deep neural networks for raga identification in hindustani classical music,” 06 2019, pp. 197–201.

[21] S. Madhusudhan and G. Chowdhary, “Deepsrgm - sequence classification and ranking in indian classical music via deep learning,” in *ISMIR*, 2019.

[22] D. P. Shah, N. M. Jagtap, P. T. Talekar, and K. Gawande, “Raga recognition in indian classical music using deep learning,” in *Artificial Intelligence in Music, Sound, Art and Design*, J. Romero, T. Martins, and N. Rodríguez-Fernández, Eds. Cham: Springer International Publishing, 2021, pp. 248–263.

[23] L. Nanni, Y. M. Costa, R. L. Aguiar, R. B. Mangolin, S. Brahnam, and C. N. Silla, “Ensemble of convolutional neural networks to improve animal audio classification,” *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2020, pp. 1–14, 12 2020. [Online]. Available: *https://asmpeurasipjournals.springeropen.com/articles/10.1186/s13636-020-00175-3*

[24] Z. Mushtaq and S.-F. Su, “Efficient classification of environmental sounds through multiple features aggregation and data enhancement techniques for spectrogram images,” *Symmetry*, vol. 12, no. 11, 2020. [Online]. Available: *https://www.mdpi.com/2073-8994/12/11/1822*