Ingénierie des Systèmes d'Information
Vol. 31, No. 4, April, 2026, pp. 1267-1272
Journal homepage: http://iieta.org/journals/isi

Automated Raga Identification System with Web Scraping Framework

Rathan Kumar Chenoori[1*], Sunil Kumar Thota[1], Suresh Lokhande[2], Srinadh Swamy Majeti[3], K Vishal Reddy[4], P. Bala Krishna[1], Pillareddy Vamsheedhar Reddy[1], Kakarlamudi V S Sudhakar[1]

[1] Department of CSE, Keshav Memorial Engineering College, Hyderabad 500088, India
[2] Department of CSE, University College of Engineering, Osmania University, Hyderabad 500007, India
[3] Department of CSE, School of Engineering, Anurag University, Hyderabad 500088, India
[4] Department of CSE, Keshav Memorial Institute of Technology, Hyderabad 500029, India

Corresponding Author Email: rathanoucse@gmail.com

https://doi.org/10.18280/isi.310421

Received: 21 July 2025
Revised: 1 February 2026
Accepted: 10 March 2026
Available online: 30 April 2026

Keywords:
Raga recognition, Indian classical music, music information retrieval, Mel-Frequency Cepstral Coefficients, Convolutional Neural Network, Bidirectional Long Short-Term Memory, attention mechanism, Explainable Artificial Intelligence

ABSTRACT

Raga recognition from audio is still challenging due to Indian classical music's intricate melodies and tonal inflections that are indicative of each Raga. Traditional methods that extract low-level simple audio features and use classifiers usually cannot capture such intricate characteristics and thus are not suitable for different performances. This work is to develop a system that can detect Hindustani Ragas and provide context with web-based information. The proposed method employs Mel-Frequency Cepstral Coefficients (MFCCs) to extract features from the audio signals and the classification of Ragas is performed using a hybrid Convolutional Neural Network–Bidirectional Long Short-Term Memory (CNN–BiLSTM) model integrated with an attention mechanism. Around 500 hours, a training dataset of 15,000 audio clips over 32 Ragas was used. Also, web scraping features were introduced to fetch data for each Raga, its history, emotional character etc. Test results indicate a classification accuracy of 92.4%, outperforming the existing deep learning models. The incorporated explainability frameworks (Grad-CAM++ and SoundLIME) allow the analysis of model predictions. This method facilitates the enhancement of music information retrieval by combining machine learning techniques and incorporating the cultural perspective of Indian classical music.

1. INTRODUCTION

With its deep cultural history of many centuries, Indian classical music is one of the world's most developed musical traditions. This tradition appears in two major styles: Hindustani Classical Music, which is found in the northern parts, and Carnatic Music, which is found in the southern parts of the country. Both traditions, however, are based on the idea of Raga - a melodic template which guides improvisation, composition and performance [1, 2], though the details differ between the two in terms of scale, ornament and emotional connotation (rasa) [3, 4].

A Raga is far more than scale or mode: it includes melodic identity, emotional flavour (rasa), catch phrases (pakad), microtonal ornamentation (gamakas) and a hierarchy of notes (vadi and samvadi) [3, 4]. These subtle aspects are what differentiate Ragas from Western tonal system, and make the Automatic Raga Identification (ARI) a very challenging problem in the computational musical analysis domain. In Western music, key identification generally involves only two steps, namely, computing a tonal center (also called the tonic or the key note), and estimating a distribution of scale degrees which are the pitch classes in the key — both can be easily computed or estimated from the distribution of notes in a song based on a small number of different heuristic algorithms, but Raga identification is much more complex, since it requires analyzing temporal patterns involving subtle pitch inflections and even aesthetic considerations that are unique to each Raga. The complications continue when one brings into the mix the various instruments and regional styles of performing, and individual artists who can take the same melodic mold and fashion infinitely different expressions [5].

1.1 Research gap and limitations of prior work

Research on automatic Raga recognition started gaining attention in the late 1990s, when most studies depended on handcrafted musical features together with conventional machine learning techniques. One of the early attempts employed Support Vector Machines (SVM) using pitch-class and dyad distribution features and reported an accuracy of nearly 82.5% while classifying 22 Ragas [3]. In a similar direction, researchers also experimented with Hidden Markov Models (HMMs) to represent pitch transitions and temporal behaviour in performances. Other classifiers such as k-Nearest Neighbors (KNNs) and Gaussian Mixture Models (GMMs) were applied using pitch histograms and Mel-Frequency Cepstral Coefficients (MFCCs) as input features for recognition tasks [6]. Although these approaches produced encouraging baseline results, their reliance on manually

designed features limited their ability to cope with the stylistic diversity that characterizes Indian classical music performances.

A further difficulty for early Raga recognition research was the limited availability of balanced annotated datasets. The widely used Raga Recognition Dataset (RRD), for instance, contains only about 116 hours of recordings covering 30 Ragas, whereas the Saraga dataset provides roughly 43.6 hours of audio spanning 61 Ragas and 9 talas [6-8]. In both datasets, frequently performed Ragas such as Yaman and Bhairavi appear much more often than several lesser-performed Ragas, resulting in noticeable class imbalance. Because of this uneven representation, models trained on these collections often show reduced performance when identifying underrepresented Ragas, which affects their overall reliability in practical applications.

In addition, several research make use of private datasets, which are not open to the public, resulting in lack of duplication and to evaluate among different works. Traditional approaches also struggled to simultaneously capture short-term musical details like ornamentations and pitch glides along with long-range melodic progression, even though accurate Raga identification requires understanding patterns that unfold across multiple time scales.

2. RELATED WORK

The recent progress in deep learning has significantly enhanced the performance of automatic Raga recognition in that it allows the models to directly learn the intrinsic complicated musical patterns from the low-levelaudio feature. Convolutional Neural Networks (CNNs) are able to effectively extract local time–frequency information from spectrograms and Mel-Frequency Cepstral Coefficient (MFCC) features, HMM based models allow to model sequential flow of melodies of Ragas, whereas Long Short-Term Memories (LSTMs) [4, 7, 9-12]. Using these properties, CNN-based methods looking at tonal and pitch contour features got up to 85% for 30 Ragas. Even better results came about when using explainable AI, reaching 88.5% for visual explanations of what the model decided.

This paper addresses these key issues. It does this by using end-to-end modeling with current deep learning architectures, organizing a large dataset, and setting up smart information retrieval. Here's what we mainly contribute:

1. Hybrid CNN-BiLSTM-Attention Architecture: We presenta new neural architecture which simultaneously performs local feature extraction via convolution layers and captures forward and backward temporal dependencies via bidirectional LSTM units, and introduces an attention mechanism which allows the model to attend to important melodic fragments. It is worth noting that this architecture takes explicitly into account the multi-scale aspect of the Raga traits, by learning not only instantaneous spectral features, but also long-term sequential patterns.

2. Large-Scale Annotated Dataset: We have assembled a dataset with 15,000 audio clips summing up to 500 hours for 32 different Ragas - which is, by far, the biggest publicly documented dataset in this domain. The dataset includes a wide range of sources, such as digital archives, streaming services, and university collections, to capture different styles of performance, artists, and recording environment. To address the problem of class imbalance, we applied the data augmentation method that includes pitch shifting, time stretching and dynamic range compression.

3. Educational Augmentation with Raga Web Scraping: because the pedagogical function of Raga recognition extends beyond the task of classification, we added an automated web scraping module to our system that assembles contextual information-historical information, performance practice, emotional connotations (rasa), and well known pieces-for a given Raga.

4. Explainable AI for Musical Performance: We utilized post-hoc interpretability techniques such as Grad-CAM++ and SoundLIME to identify the audio characteristics that affect the model's prediction to provide better insight into the results.

5. Cross-Dataset Evaluation: We perform evaluations on other datasets, including Saraga and YouTube recordings, to verify the ability of the proposed approach to generalize well to multiple recording settings.

3. DATASET

The performance of any ARI system is highly reliant on the quality of the training data [13-16]. However, the popular Hindustani Classical Music Datasets (HCMDs) namely RRD and Saraga suffer with a number of limitations hindering dnn based research. For instance RRD comprises of ca. 116 hours of recordings for 30 Ragas and is unbalanced across classes, while Saraga is imbalanced as well and in addition class overlapping problem and very few samples for some Ragas.

In this paper, we address this challenge by presenting a significantly larger and more balanced dataset as shown in Table 1 with 15,000 labeled audio clips aggregating to almost 500 hours for 32 Ragas as shown in Table 1.

Table 1. Description of dataset

Attribute	Value
Total Audio Clips	15,000
Total Duration	500 hours
Raga Classes	32
Avg. Duration / Clip	~2 minutes

4. METHODOLOGY

4.1 Feature extraction

To identify the timbral and spectral quantities of music used MFCCs as:

1) Pre-Emphasis

A pre-emphasis filter is applied to boost high-frequency components and improve signal-to-noise ratio:

$$y[n] = x[n] - \alpha . x[n-1] \tag{1}$$

where, α is typically set to 0.97.

2) Framing and Windowing

If N is the frame size, the window function is:

$$w[n] = 0.54 - 0.46 . \cos\left(\frac{2\pi n}{N-1}\right), 0 \le n < N \tag{2}$$

3) Short-Time Fourier Transform (STFT)

$$P[k] = \frac{1}{N}|X[k]|^2 \tag{3}$$

where, X[k] is the discrete Fourier transform of the frame.

4) Mel Filter Bank

A set of triangular filters spaced on the Mel scale is applied to the power spectrum. The Mel scale converts linear frequency f (in Hz) to the Mel scale via:

$$m = 2595.log_{10}\left(1 + \frac{f}{700}\right) \quad (4)$$

5) Logarithm and Discrete Cosine Transform (DCT):

$$c_n = \sum_{m=1}^{M} log(E[m]).cos\left[\frac{\pi n(m-0.5)}{M}\right] \quad (5)$$

M is the number of Mel filters [17, 18].

4.2 Deep neural network architecture

The extracted MFCC sequences are fed into the hybrid deep learning architecture as shown in Figure 1:

- The CNN block identifies local time–frequency patterns, such as characteristic note transitions and ornamentations (*gamakas*).
- The BiLSTM layer models temporal dependencies in both forward and backward directions.
- The Attention mechanism dynamically assigns higher weights to the most informative time segments.

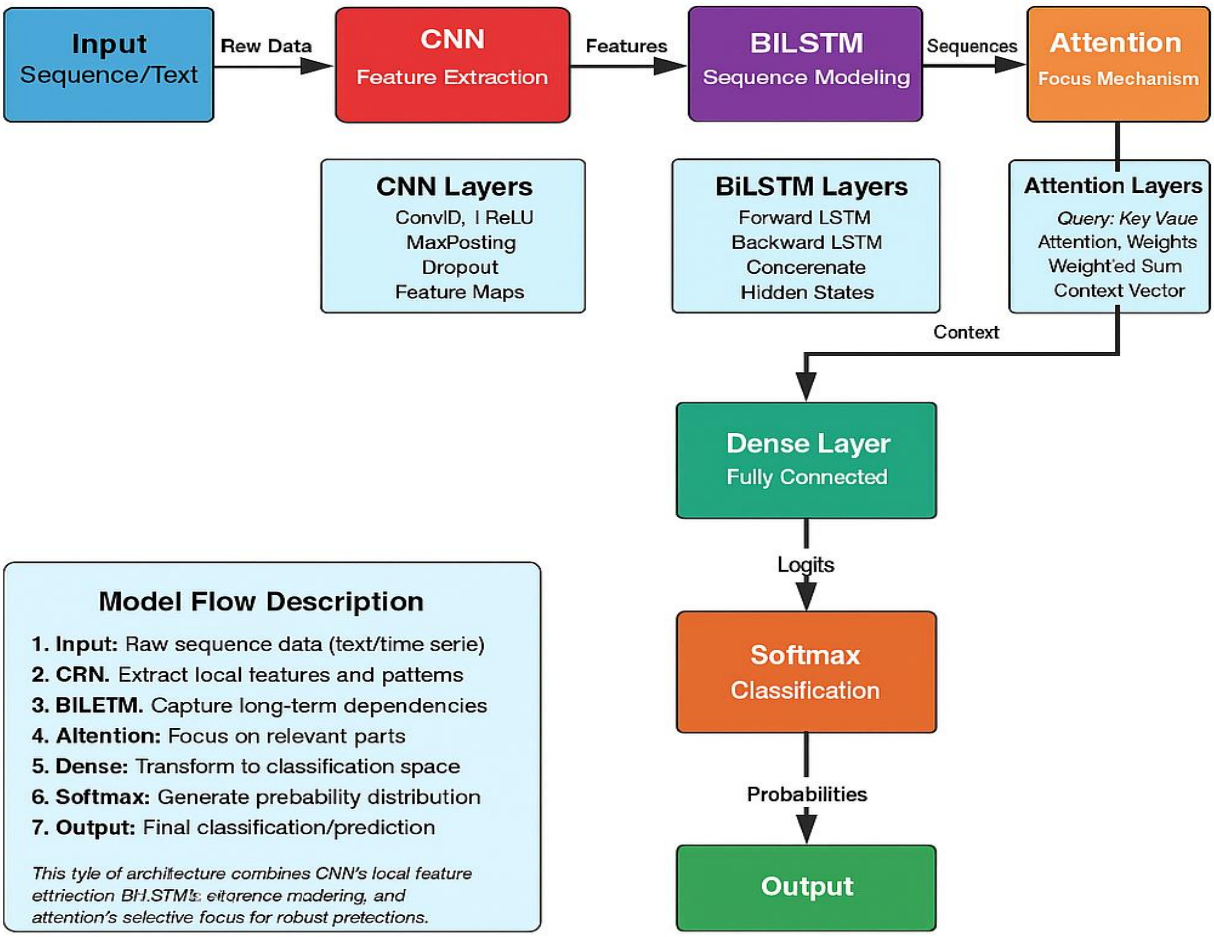

Figure 1. Architecture diagram of CNN + BiLSTM + Attention

Then the attention mechanism assigns weights α_t to each time step t, computed as:

$$\alpha_t = V^T \tan h(W_h h_t + b_h) \quad (6)$$

where, h_t denotes the hidden state at time t, and V, W_h, b_h are learnable parameters.

4.3 Model training

The training process was executed with a Adam optimizer, a learning rate of 1×10^{-3} and the categorical cross-entropy loss function. Reserve dense layers were regularized by dropout to avoid overfitting. We assessed the model performance with stratified 10-fold cross-validation, which ensures robust performance estimation for all Ragas [19].

5. EXPERIMENTAL RESULTS AND DISCUSSION

5.1 Cross-dataset evaluation

About 60% of the recordings were from Saraga and CompMusic archives, which contain professionally annotated concert recordings with metadata such as raga, tala, and artist [20-22]. To train the model on data from beyond these two

sources, supplementary audio files were sourced from YouTube channels dedicated to classical music, archival performances of All India Radio, as well as from cloud-based repositories such as SoundCloud and Sangeetpedia, among others, give particular attention to those tagged explicitly with raga names. A manual check was performed for every downloaded file, the identity of the raga and the tonal correctness (through the characteristic phrases or pakad and through the note progression in the scale or the aroha–avaroha) was verified independently by two trained musicians. To ensure that the labels were not corrupted, unlabeled or controversially labeled recordings were discarded.

Table 2. Performance across test sets

Dataset	Chunk-Wise F1-Score (%)	Song-Wise F1-Score (%)
Internal Test Set (St)	92.4	94.1
Saraga Dataset	89.7	91.0
YouTube Samples	88.3	90.2

For Saraga, 23 audios were chosen for 11 Ragas among 32 Ragas considered in this study. The Nayaki-Kanada is not found in Saraga we hence keep the examples from St for this class. Sample distribution in Saraga Bhairavi has 5 files, for other Ragas have less than three examples each. We also select 55 recordings from YouTube where each raga class is guaranteed to be represented by at least four and at most five performances. All audio is preprocessed, as described in Section IV, including normalization, silence removal and extraction of MFCC features. To: For tonic normalization, Saraga's metadata has a direct access to the tonic value. Tonic estimation for YouTube samples is based on the CompIAM-tk toolkit introduced in the study [12] achieving accuracy = 91.76% on our internal validation. Table 2 presents the performance of the proposed system across different test datasets.

Evaluation metrics include:

- **Chunk-wise F1-score** computed across 30-second segments
- **Song-wise F1-score**, determined by majority voting across segments of each song

5.2 Saliency-based evaluation

Beyond classification accuracy, we analyze the explainability of our model through two XAI methods: Grad-CAM++ (GC) and SoundLIME (SL). These analyses aim to verify whether the model bases predictions on musically meaningful audio segments.

1. Expert Saliency Annotation

We gather a subset of 64 30-second audio clips that are correctly classified and are randomly sampled from St. A HCM domain expert annotates segments with signature musical formations such as Pakad, Aaroh, Avroh, or Mukh-Ang. Only 5–7 and 15–17 s long segments are labeled, so no obvious or overly wide segments are selected. The final saliency labeled test set S_{ts} has 47 examples from all 12 Ragas with a minimum of 3 examples per class.

2. Saliency Evaluation Methodology

For Grad-CAM++, the time-frequency saliency maps are reduced along the frequency axis to produce a 1D saliency vector highlighting important time frames. For SoundLIME, local importance weights are derived from a linear surrogate model fitted to perturbed spectrogram regions, subsequently interpolated to align with expert annotation timings.

Table 3 presents precision scores over time, showing SoundLIME's consistent advantage over Grad-CAM++, especially for shorter periods.

3. Interpretability-Performance Relationship

Table 4 gives specific instances that compare expected and real classes, model certainty, and saliency exactness for both XAI methods.

Table 3. Saliency precision at top t seconds

Top t Seconds	Precision – GC	Precision – SL
1	0.64	0.64
2	0.58	0.58
3	0.52	0.52
5	0.46	0.46
10	0.39	0.39

Note: GC = Grad-CAM; SL = SoundLIME.

Table 4. Example saliency analysis

Sample	True Class	Predicted Class	Softmax Probability	GC Precision	SL Precision
Des	Des	Des	0.95	0.77	0.80
Bihag	Bihag	Maru-Bihag	0.71	0.12	0.00

Table 5. Comparison table

Study / System	Feature Set	Model Type	Dataset Used	Accuracy (%)
[3]	Pitch-class, dyad distributions	SVM	Private	~82.5
[6]	Tonal features, pitch contours	CNN	Raga Recognition Dataset	~85.0
[7]	MFCC, chromagram, LPC	KNN, GMM, HMM	Mixed datasets	70–85
[8]	MFCC, Spectrograms	CNN + LSTM (XAI applied)	ISMIR dataset subset	~88.5
Proposed System	MFCC, Derivatives	CNN + BiLSTM + Attention	Custom (500 hrs,32 Ragas)	92.4

Note: SVM = Support Vector Machine; CNN = Convolutional Neural Network; KNN = k-Nearest Neighbors; GMM = Gaussian Mixture Model; HMM = Hidden Markov Model; LSTM = Long Short-Term Memory; MFCC = Mel-Frequency Cepstral Coefficient; LPC = Linear Predictive Coding; BiLSTM = Bidirectional Long Short-Term Memory; XAI = Explainable Artificial Intelligence.

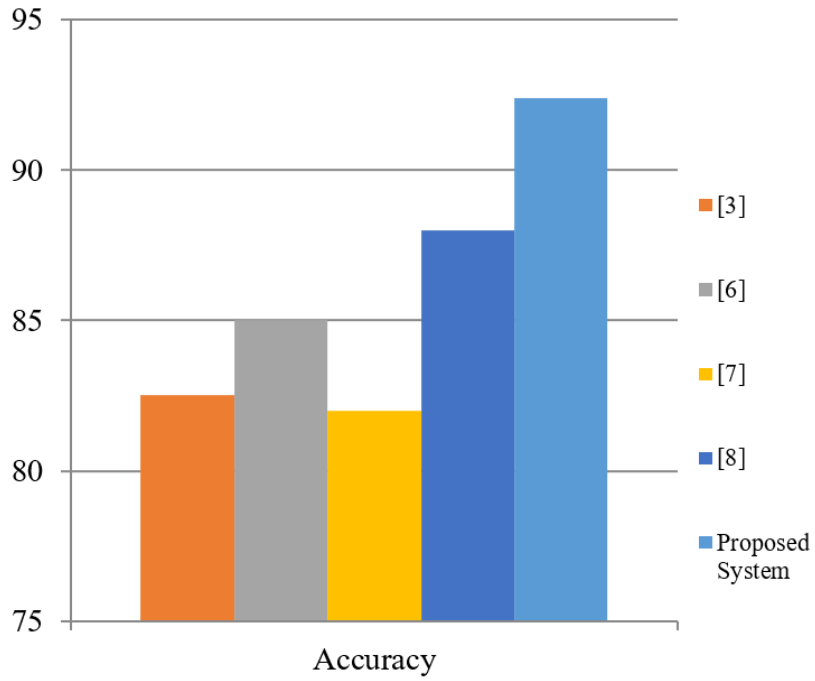

Figure 2. Accuracy comparison

Hybrid models with MFCC and chromagram features showed moderate performance across mixed datasets. The proposed CNN–BiLSTM–Attention model utilizing MFCCs and their derivatives on a 500-hour custom dataset of 32 Ragas achieved the highest accuracy of 92.4% as showned in Table 5 and Figure 2.

6. CONCLUSION AND FUTURE SCOPE

We have designed an automated system to identify the Raga of the query music track. It is a hybrid model with CNN, BiLSTM and attention mechanisms and we trained it with large dataset of Hindustani Classical Music– 15,000 clips for 32 Ragas. The intuition lies in the fact that we want to capture not only the short-term dynamics in sound, but also how the music evolves on the longer term, both of which are important for Ragas. 4 for the testing. One cool thing is that it will also automatically pull up info on each raga (history, what feelings you are supposed to have, etc.) so it is actually useful for learning. We also employed a couple of magic tools (Grad-CAM++ and SoundLIME) to visualize on which the model bases its decision.

REFERENCES

[1] Singh, P., Arora, V. (2025). Explainable deep learning analysis for raga identification in Indian art music. IEEE Transactions on Audio, Speech and Language Processing, 33: 2302-2311. https://doi.org/10.1109/TASLPRO.2025.3574839

[2] Madhusudhan, S.T., Chowdhary, G. (2024). DeepSRGM--sequence classification and ranking in Indian classical music with deep learning. arXiv preprint arXiv:2402.10168. https://doi.org/10.48550/arXiv.2402.10168

[3] Alekh, S. (2017). Automatic Raga recognition in Hindustani classical music. arXiv preprint arXiv:1708.02322. https://doi.org/10.48550/arXiv.1708.02322

[4] Srinivasamurthy, A., Gulati, S., Repetto, R.C., Serra, X. (2021). Saraga: Open datasets for research on indian art music. Empirical Musicology Review, 16(1): 85-98. https://doi.org/10.18061/emr.v16i1.7641

[5] Hebbar, D., Jagtap, V. (2022). A comparison of audio preprocessing techniques and deep learning algorithms for raga recognition. arXiv preprint arXiv:2212.05335. https://doi.org/10.48550/arXiv.2212.05335

[6] Aswale, S.P., Patil, W., Chavate, S. (2025). Raga net: A novel deep learning framework for Indian raga recognition based on deep convolution neural network and long short-term memory. International Journal of Engineering Trends and Technology, 73(2): 155-165. https://doi.org/10.14445/22315381/IJETT-V73I2P113

[7] Clayton, M., Li, J., Clarke, A., Weinzierl, M. (2023). Hindustani raga and singer classification using 2D and 3D pose estimation from video recordings. Journal of New Music Research, 52(4): 285-300. https://doi.org/10.1080/09298215.2024.2331788

[8] Bhangale, K.B., Kothandaraman, M. (2022). Survey of deep learning paradigms for speech processing. Wireless Personal Communications, 125(2): 1913-1949. https://doi.org/10.1007/s11277-022-09640-y

[9] Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, pp. 839-847. https://doi.org/10.1109/WACV.2018.00097

[10] Ribeiro, M.T., Singh, S., Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California USA, pp. 1135-1144. https://doi.org/10.1145/2939672.2939778

[11] Sridhar, R., Geetha, T.V. (2006). Swara indentification

for South Indian classical music. In 9th International Conference on Information Technology (ICIT'06), Bhubaneswar, India, pp. 143-144. https://doi.org/10.1109/ICIT.2006.83

[12] SamsekaiManjabhat, S., Koolagudi, S.G., Rao, K.S., Ramteke, P.B. (2017). Raga and tonic identification in carnatic music. Journal of New Music Research, 46(3): 229-245. https://doi.org/10.1080/09298215.2017.1330351

[13] Shetty, S., Hegde, S. (2019). Automatic classification of carnatic music instruments using MFCC and LPC. Data Management, Analytics and Innovation: Proceedings of ICDMAI 2019, 1: 463-474. https://doi.org/10.1007/978-981-32-9949-8_32

[14] Gunawan, A.A., Suhartono, D. (2019). Music recommender system based on genre using convolutional recurrent neural networks. Procedia Computer Science, 157: 99-109. https://doi.org/10.1016/j.procs.2019.08.146

[15] Elbir, A., Aydin, N. (2020). Music genre classification and music recommendation by using deep learning. Electronics Letters, 56(12): 627-629. https://doi.org/10.1049/el.2019.4202

[16] Kim, H.G., Kim, G.Y., Kim, J.Y. (2019). Music recommendation system using human activity recognition from accelerometer data. IEEE Transactions on Consumer Electronics, 65(3): 349-358. https://doi.org/10.1109/TCE.2019.2924177

[17] Tzanetakis, G., Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10(5): 293-302. https://doi.org/10.1109/TSA.2002.800560

[18] Oganian, Y., Bhaya-Grossman, I., Johnson, K., Chang, E.F. (2023). Vowel and formant representation in the human auditory speech cortex. Neuron, 111(13): 2105-2118. https://doi.org/10.1016/j.neuron.2023.04.004

[19] Dietrich, B.J., Hayes, M., O'brien, D.Z. (2019). Pitch perfect: Vocal pitch and the emotional intensity of congressional speech. American Political Science Review, 113(4): 941-962. https://doi.org/10.1017/S0003055419000467

[20] Bhangale, K.B., Kothandaraman, M. (2023). Speech emotion recognition using the novel PEmoNet (Parallel Emotion Network). Applied Acoustics, 212: 109613. https://doi.org/10.1016/j.apacoust.2023.109613

[21] Pradhan, A., Yajnik, A. (2024). Parts-of-speech tagging of Nepali texts with Bidirectional LSTM, Conditional Random Fields and HMM. Multimedia Tools and Applications, 83(4): 9893-9909. https://doi.org/10.1007/s11042-023-15679-1

[22] Korkmaz, Y., Boyacı, A. (2023). Hybrid voice activity detection system based on LSTM and auditory speech features. Biomedical Signal Processing and Control, 80: 104408. https://doi.org/10.1016/j.bspc.2022.104408