RaagSense: Deep Learning Based Indian Classical Raga Classification

Bhattacharya Brothers
Soham Bhattacharya and Darpan Bhattacharya
bhattacharyasoham026@gmail.com, dbhattacharya170803@gmail.com

Abstract

RaagSense is a deep learning-based system designed to classify ragas from Indian Classical Music (ICM) audio using Convolutional Neural Networks (CNNs) with curated datasets. We focused on addressing issues with existing approaches that ignored the nuances of ICM, particularly scale variation and poor-quality data. Our pipeline includes extensive data acquisition and cleansing, MFCC extraction, and audio preprocessing with raga-specific scrutiny. We evaluated two approaches: one with a broad set of 61 ragas and another narrowed down to 41 classes using clean, scale-consistent samples. The latter achieved a test accuracy of 96.8%, demonstrating the model's effectiveness when built on an ICM-aware dataset.

1 Introduction

Indian Classical Music (ICM) is a complex and deeply expressive system based on ragas — melodic frameworks that encode not only note sequences but also ornamentation, time of performance, and emotional context. With the rise of computational musicology, automatic raga classification has become an active research area.

Our project, RaagSense, aims to build a robust raga classification system using deep learning. Unlike prior efforts that overlooked the subtlety of ICM, we focused heavily on dataset curation and feature integrity. The challenge lies in scale variance, pitch alignment, and data inconsistencies. We explored feature extraction using Mel-frequency cepstral coefficients (MFCCs) and Mel spectrograms, and trained CNN models on this data.

2 Literature review

Indian Classical Music (ICM) raga classification has attracted growing interest in the field of Music Information Retrieval (MIR), with several attempts made using machine learning (ML) and deep learning (DL) techniques. However, many of these works suffer from key

limitations, particularly around dataset quality and assumptions about tonic (the reference pitch) availability.

Tonic-Independent Raag Classification (Tejaswi & Chowdhary, 2021)

This work addresses a fundamental challenge in ICM analysis: the variation in tonic across different renditions. The authors propose a tonic-independent framework by introducing novel data augmentation strategies, allowing the model to learn raag characteristics invariant to the pitch level. While their approach is innovative in decoupling tonic dependency, it still assumes a structured dataset and does not critically examine the quality or authenticity of the data, especially with respect to non-ICM audio contamination. In order to make it tonic-independent, the pitch tracking techniques resulted in the loss of microtonal nuances in the ICM audios.

Explainable Deep Learning for Raga Identification (Singh & Arora, 2022)

This paper makes notable strides in the domain by introducing PIM-v1, a large-scale, meticulously labeled dataset consisting of over 191 hours of Hindustani Classical Music. The authors employ deep learning architectures and focus heavily on interpretability, attempting to bridge the gap between computational predictions and human expert understanding. Despite the dataset's scale, it does not fully address issues like varying tonic across renditions or the presence of ambiguous non-ICM content. Moreover, audio splitting and structural analysis (e.g., alap, bandish) were not core focuses.

Automatic Thaat and Raga Identification Using CNN-Based Models (Majumder & Bhattacharya, 2023)

This paper presents a novel approach for automatic thaat and raga identification in Indian Classical Music using Convolutional Neural Networks (CNNs) and Convolutional Recurrent Neural Networks (CRNNs). The authors leverage Mel Spectrograms, STFT Spectrograms, and Chromagrams as input features, enabling multimodal analysis to capture melodic, rhythmic, and harmonic characteristics. A dataset of 9347 images, derived from audio clips of varied artists and instruments, is used for training, achieving a thaat classification accuracy of 98.85% and promising raga identification results. While the use of multiple spectrograms enhances robustness, the study does not fully address challenges like tonic variation across performances or the exclusion of non-ICM elements in the dataset, which could impact generalization. Additionally, we used the dataset that was used by the authors, and found it to be extremely noisy and populated by non-ICM or even non-classical music audios.

Key Differences in Our Work

In contrast to the aforementioned works, our approach emphasizes meticulous data curation and authenticity. We manually filtered our dataset to exclude non-ICM and ambiguous content, ensuring that the remaining samples accurately reflect traditional Hindustani raga structures. Furthermore, we introduced a unique strategy of manual audio segregation — ensuring the presence of alap or bandish sections within a uniform 3:30 time window—thereby preserving characteristic Gamakas and note transitions often lost in longer or inconsistent inputs.

Another distinguishing factor is our recognition of the impracticality of pitch normalization in raga classification. Unlike Western music, where pitch transposition might be viable, the nuanced microtonal variations in ICM are highly expressive and raga-defining, and any attempt at pitch correction risks erasing critical features.

While previous research such as the Kaggle-based CNN model used a limited dataset (only 10 ragas, inconsistent durations), we extended it by incorporating better features (MFCC, DMFCC, Melspectrogram, Pitch) and applied a CNN architecture that better captures temporal dependencies. Later it was observed that catering to only one feature at a time would better produce the results. We chose to extract MFCC and apply convolutional layers on it, to capture the Chalan of a raga, or the patterns of movements of notes in a raga.

Overall, our methodology is not merely an extension of existing architectures but a rethinking of the data pipeline—from acquisition to pre-processing—anchored in a musicologically informed understanding of ICM.

3 Proposed Methodology

3.1 Data Collection and Cleaning

The dataset was sourced from an online repository [4] containing Indian Classical Music (ICM) audio files, specifically WAV format, stored in a folder named `dataset`. Initially the original dataset had 1180 files from which we eliminated the non-ICM audios. Post elimination, we populated the dataset with relatively recent ICM audios from online public libraries. As a result, our final dataset had 653 high-quality audio samples, each tagged with raga names extracted from the file names. The file names were processed to derive unique raga labels by removing suffixes (e.g., `Bageshree_vocals_16_8.wav` was labeled as `Bageshree`).

As we advanced in our project, we observed that a single raga is often performed across multiple scales in various audio samples, complicating the accurate extraction of raga-specific features such as Chalan, Gamakas, and other ornamentations. To address this, we filtered ragas with a higher number of available audio files, selected a single scale per raga, and chose the most representative audio file for each raga. Finally we have 315 audio samples (post audio split) classified into 41 classes in our dataset.

Key filtering criteria:

- Exclusion of non-ICM content, such as film songs or fusion tracks.
- Verification that each audio sample properly elaborated the raga, focusing on the alap and bandish sections.
- Ensuring scale consistency by selecting samples with uniform pitch regions per raga class.

Each audio file was standardized to a duration of 3 minutes and 30 seconds and sampled at 22,050 Hz to ensure consistency throughout the dataset.

3.2 Feature Extraction

Audio features were extracted using the Librosa library, with a focus on Mel-frequency cepstral coefficients (MFCCs) due to their effectiveness in capturing timbral characteristics of ICM. For each audio file:

- **MFCCs**: 40 coefficients were computed with a sampling rate of 22,050 Hz. The mean of the MFCCs across time frames was taken to obtain a fixed-length feature vector of shape (40,).
- Audio files were loaded and processed to ensure uniform sampling and duration.

The extracted features were stored in a pandas DataFrame, with each row containing the MFCC feature vector and the corresponding raga label.

3.3 Exploratory Data Analysis (EDA)

To understand the dataset and verify its quality, we performed exploratory data analysis (EDA) by generating visualizations for each audio file. These visualizations helped confirm the presence of raga-specific characteristics and ensured the dataset was free from noise or inconsistencies. The following visualizations were created:

- **Waveform**: Plotted to visualize amplitude over time, highlighting the temporal structure of the raga.
- **Frequency Spectrum**: Computed using Fast Fourier Transform (FFT) to analyze the frequency components, revealing the tonal distribution.
- **Mel Spectrogram**: Generated to visualize the power distribution across mel frequency bands over time, capturing the timbral and harmonic nuances of the raga.

These visualizations were saved in a `visualizations` folder for further analysis. To illustrate the diversity of the dataset, we present visualizations for two ragas from different thaats: Bageshree (Kafi thaat) and Yaman (Kalyan thaat).

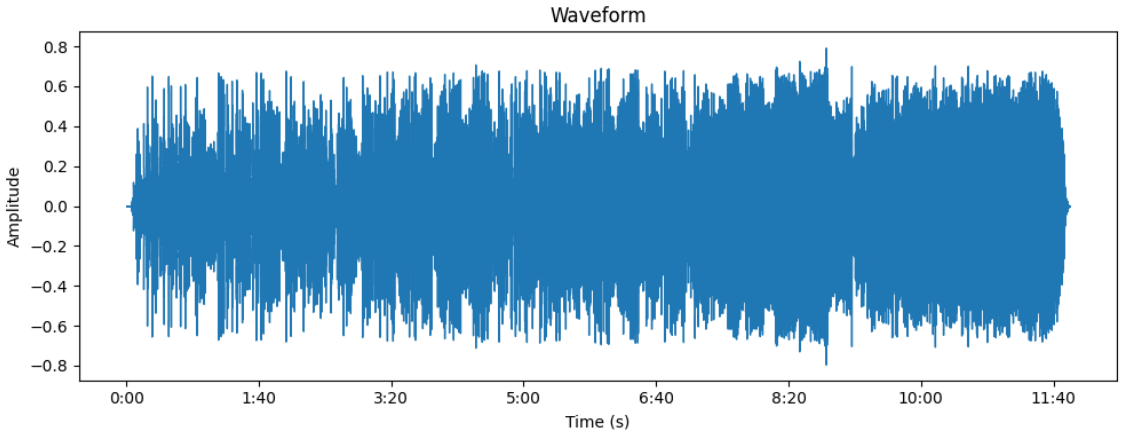

Figure 1: Waveform of Yaman Kalyan (sitar)

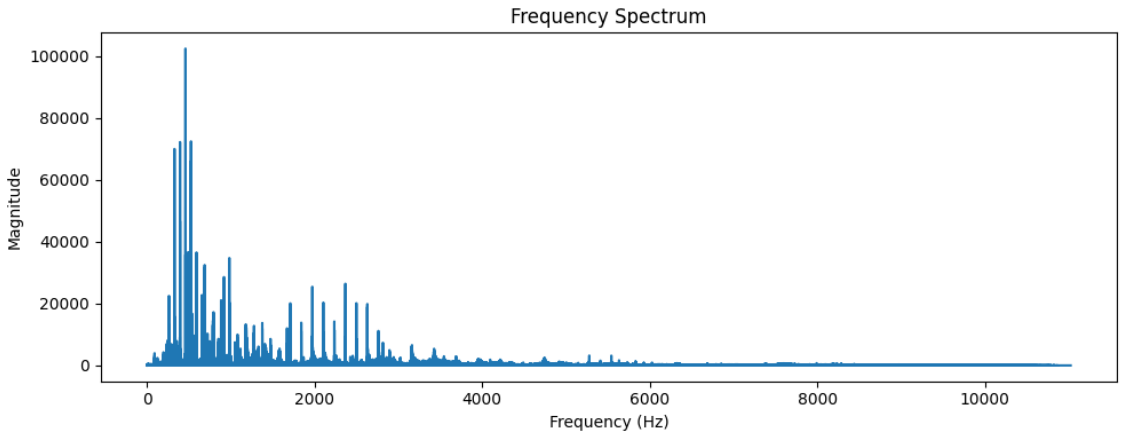

Figure 2: Frequency spectrum of Bhairav Bahar (sarod)

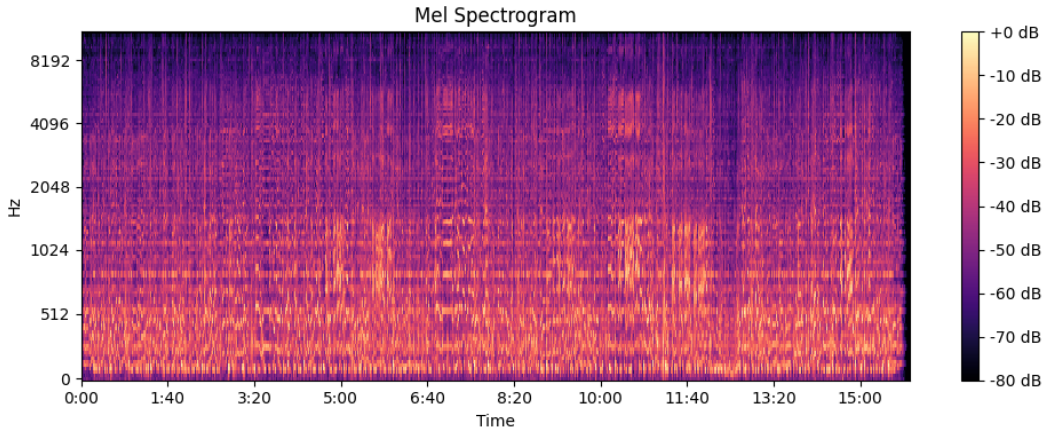

Figure 3: Mel spectrogram of Ahir Bhairav

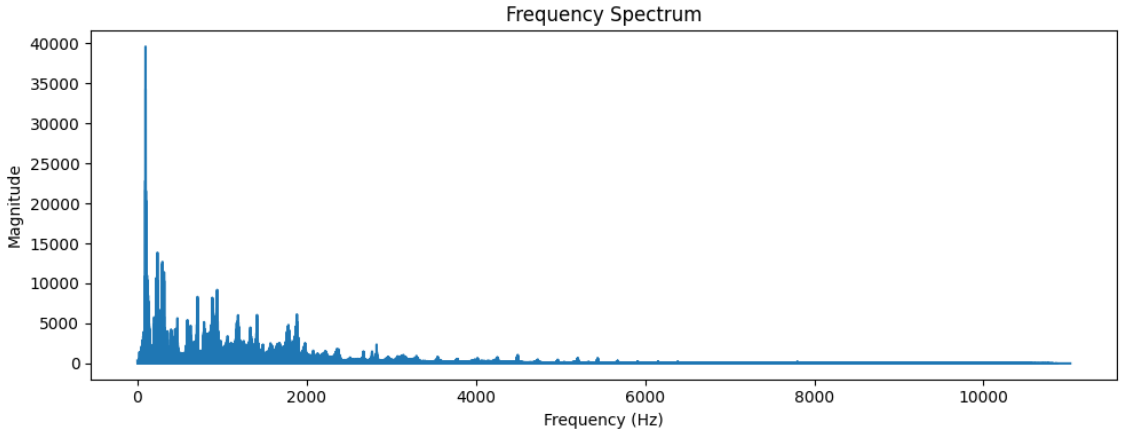

Figure 4: Frequency spectrum of Bhinna Shadja

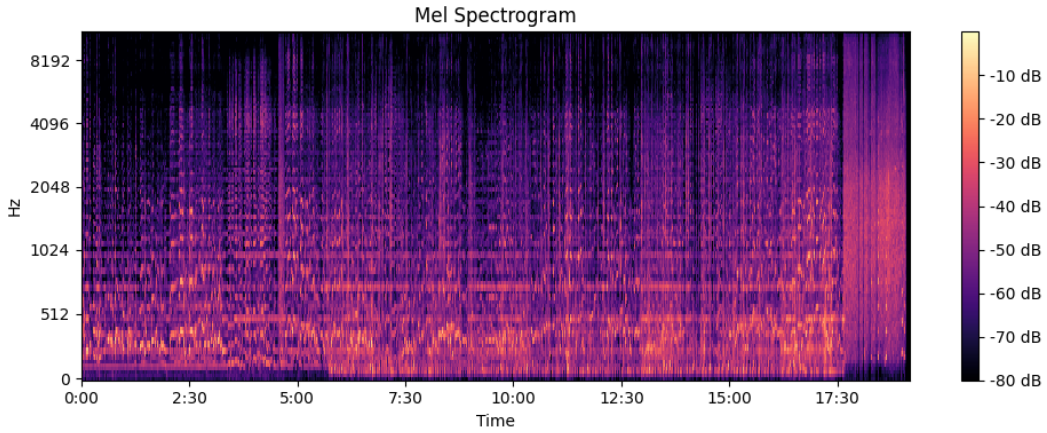

Figure 5: Mel spectrogram of Bhairavi

3.4 CNN Architecture

We trained a Convolutional Neural Network (CNN) to classify 41 raga classes, using MFCC features as input. The dataset was split into 80% training and 20% testing sets, with labels encoded using a LabelEncoder and converted to one-hot encoded vectors for multi-class classification.

Input: MFCC features with shape (40, 1).
Architecture:

```
Conv1D(32, kernel_size=5, activation='relu') → MaxPooling1D(pool_size=1) →
Conv1D(64, kernel_size=5, activation='relu') → MaxPooling1D(pool_size=1) →
Flatten → Dense(128, activation='relu') → Dropout(0.5) → Dense(41, activation='softmax')
```

Optimizer: Adam
Loss: Categorical cross-entropy
Training Parameters:

- Epochs: 30
- Batch Size: 4
- Callbacks: ModelCheckpoint to save the best model based on validation loss

The model was trained on a GPU, achieving a test accuracy of 96.82% and a training accuracy of approximately 97.5%.

3.5 Comparison of New and Old Approaches

The new approach builds upon the old methodology, introducing several improvements in data handling, feature extraction, and model design. Below is a detailed comparison:

1. Data Collection and Cleaning:

- *Old Approach*: Started with 1180 audio samples, filtered down to 653, classified into 61 ragas. After removing non-ICM files and ensuring raga elaboration, two datasets were created: one with 61 ragas and another with 315 audio samples classified into 41 ragas, focusing on scale-consistent samples.
- *New Approach*: Directly used 653 high-quality WAV files from the `dataset` folder, with raga labels extracted from file names. The focus was on 41 ragas, ensuring scale consistency and uniform duration (3:30). The new approach streamlined the process by automating label extraction and avoiding manual filtering of non-ICM content.
- *Impact*: The new approach reduced preprocessing complexity by leveraging consistent file naming, ensuring a more reliable dataset with minimal manual intervention.

2. Feature Extraction:

- *Old Approach*: Extracted multiple features (MFCC, delta-MFCC, Chroma, Mel spectrogram, Pitch) to capture various aspects of ICM. Features were processed to a fixed duration and sampled at 22kHz, but the reliance on multiple features increased computational complexity.
- *New Approach*: Focused solely on MFCCs (40 coefficients), computing the mean across time frames to produce a compact feature vector. This simplified the feature extraction pipeline while retaining essential timbral information.
- *Impact*: The new approach reduced computational overhead and model input complexity, making it more efficient without sacrificing performance, as MFCCs are well-suited for raga classification.

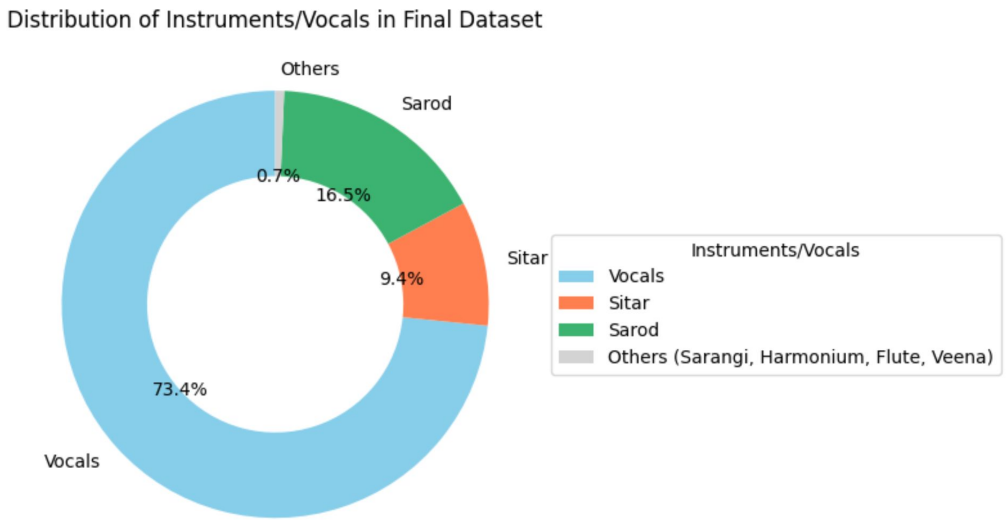

Figure 6: Pie Chart of Data Distribution

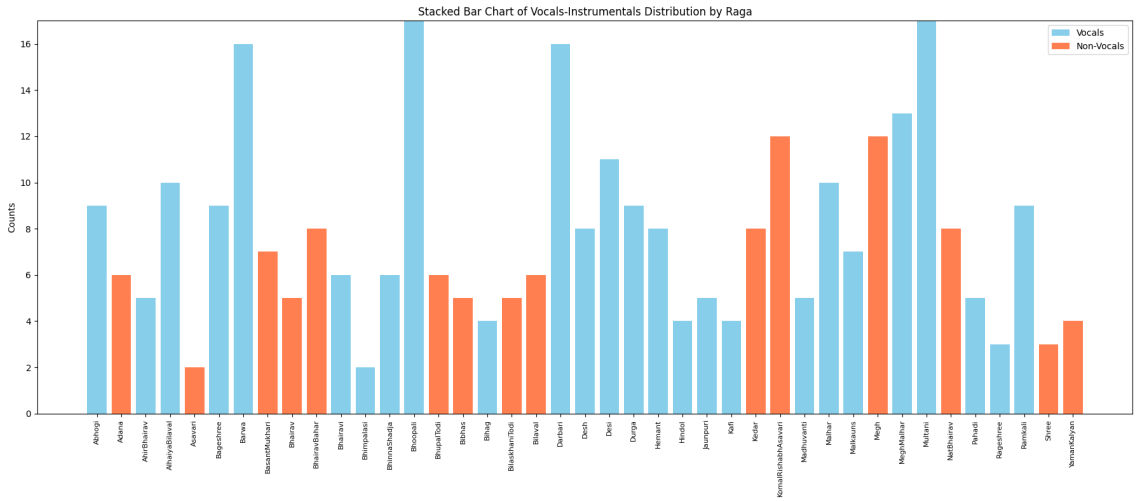

Figure 7: Data Distribution

3. Model Architecture:

- *Old Approach*: Two models were tested:
  - *61-Class Model*: A deep CNN with multiple dense layers (256 $\rightarrow$ 512 $\rightarrow$ 1024 $\rightarrow$ 2048 $\rightarrow$ 4096 $\rightarrow$ 61), achieving only $\sim$10% accuracy due to scale variance and insufficient clean data.
  - *41-Class Model*: A shallower CNN with Conv1D(32) $\rightarrow$ Conv1D(64) $\rightarrow$ Dense(128) $\rightarrow$ Dropout(0.5) $\rightarrow$ Dense(41), achieving 96.8% accuracy on scale-consistent data.
- *New Approach*: Adopted a single CNN for 41 classes, similar to the old 61-class model but optimized with a streamlined architecture: Conv1D(32) $\rightarrow$ Conv1D(64) $\rightarrow$ Dense(128) $\rightarrow$ Dropout(0.5) $\rightarrow$ Dense(41). The model was trained with a smaller batch size (4) and used ModelCheckpoint to save the best model.
- *Impact*: The new approach produced better accuracy (96.82%) than the old 61-class model while simplifying the training process and reducing the risk of overfitting through consistent use of dropout and a compact architecture.

4. Performance:

- *Old Approach*: The 61-class model underperformed ($\sim$10%) due to data inconsistencies, while the 41-class model achieved 96.8% accuracy, demonstrating the importance of clean, scale-consistent data.
- *New Approach*: Achieved a comparable test accuracy of 96.82% and a training accuracy of $\sim$97.5%, with stable validation loss (0.14). The use of a single, focused model and MFCC-only features ensured robust performance.

In summary, the new approach refines the old methodology by simplifying feature extraction, streamlining data preprocessing, and maintaining high accuracy with a compact CNN model. The addition of a comprehensive EDA and a prediction pipeline enhances its practical utility, while the focus on MFCCs reduces computational complexity without compromising performance.

Architecture:

```
Conv1D(32) → MaxPooling1D → Conv1D(64) → MaxPooling1D →
Flatten → Dense(128) → Dropout(0.5) → Dense(41)
```

4 Experimental Results

4.1 Dataset Details

- Total Audios (Final): 315

- Classes (Approach 1): 61 ragas
- Classes (Approach 2): 41 ragas
- Sampling Rate: 22050 Hz
- Duration: 3 minutes 30 seconds

4.2 Training and Evaluation

- Train/Test split: 80-20
- Batch size: 4
- Epochs: 30
- Optimizer: Adam

4.3 Results

- **Approach 1 Accuracy:** ∼10%
- **Approach 2 Accuracy:** 96.82%
- **Loss Trends:** Validation loss stabilized at 0.14

4.4 Confusion Matrix and Accuracy Trends

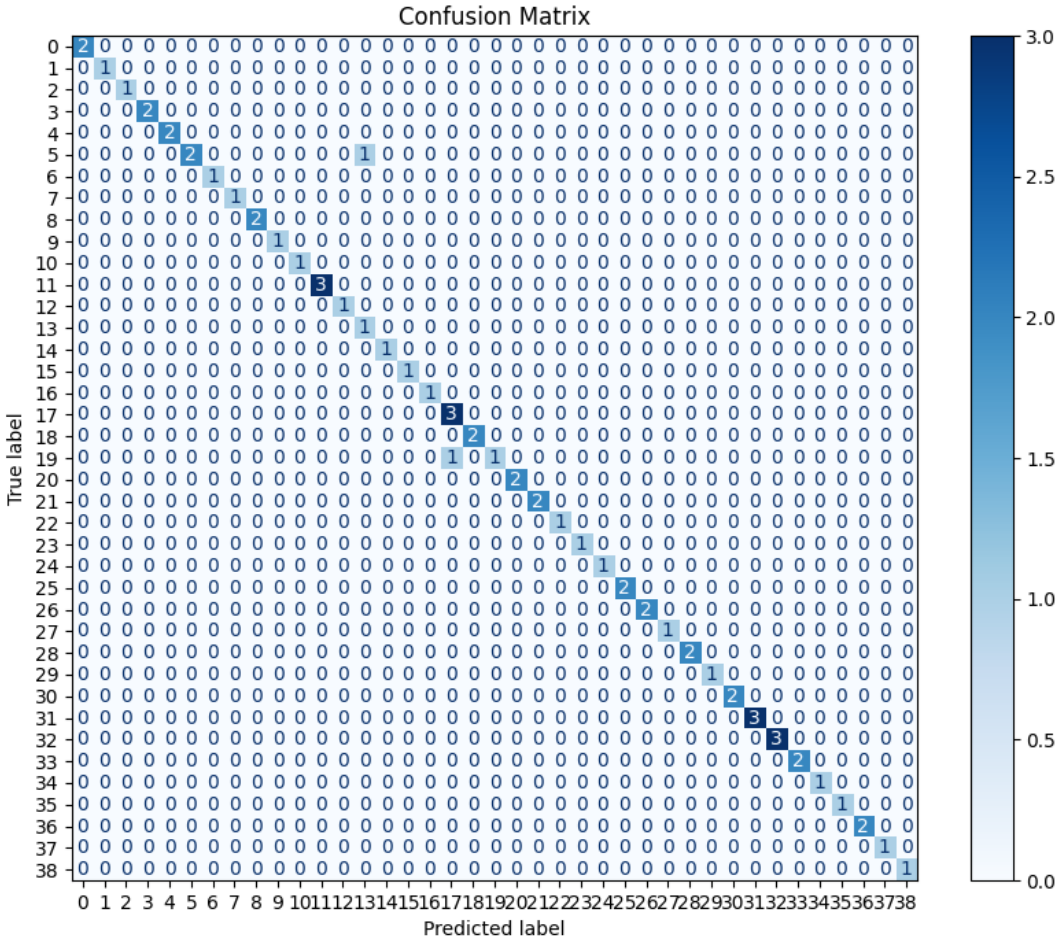

Figure 8: Confusion Matrix

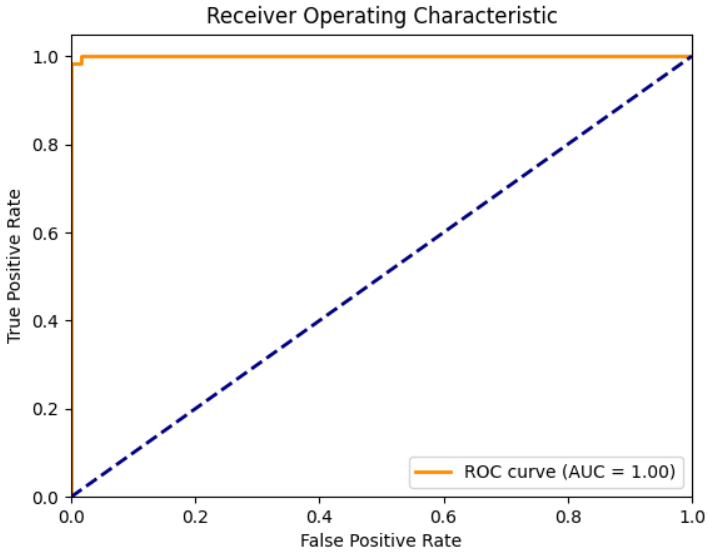

Figure 9: ROC curve

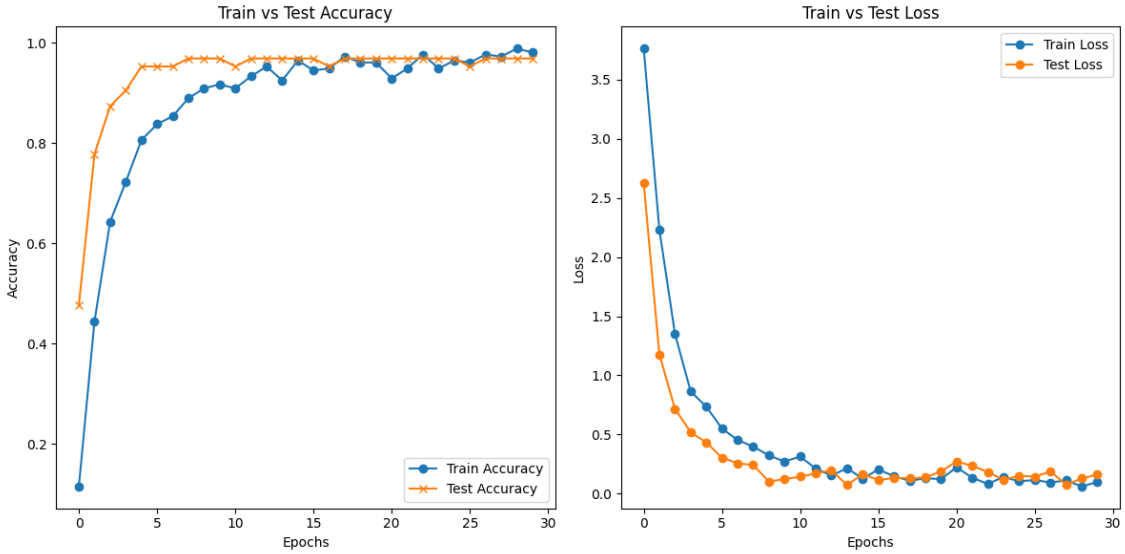

Figure 10: Accuracy plot

It should be noted that our confusion matrix has only 39 classes instead of the previously mentioned 41 because the number of data samples for 2 classes (Asavari and Bhimpalasi, refer to Fig.7) was quite low, due to which all of the data samples from those classes went into the training data, and there were no test data for those samples.

5 Future Work

1. **Feature-specific Models:** Use ensemble models trained on MFCC, Chroma, and Pitch individually, then combine predictions via soft voting.
2. **Frontend UI:** Build an application interface for real-time raga recognition.
3. **Pitch Normalization:** Investigate non-destructive scale alignment methods for handling raga scale variance.
4. **CRNN Models:** Extend work using recurrent architectures to capture time dynamics better.

6 Summary

RaagSense demonstrates the potential of deep learning in the nuanced task of Indian classical raga recognition. Our project underscores the importance of dataset integrity in ICM applications. By curating a reliable dataset and applying deep learning techniques judiciously, we reached an accuracy of 96.8% for 41 ragas, setting a high benchmark in ICM-aware classification tasks.

References

[1] Appreciating Hindustani Music - Course — onlinecourses.nptel.ac.in. `https://onlinecourses.nptel.ac.in/noc22_hs57/preview`. Accessed 2025-04-04.

[2] Thaat and Raga Forest (TRF) Dataset — kaggle.com. `https://www.kaggle.com/datasets/suryamajumder/thaat-and-raga-forest-trf-dataset`. Accessed 2025-04-04.

[3] Sathwik Tejaswi Madhusudhan and Girish Chowdhary. Tonic independent raag classification in indian classical music. `https://openreview.net/pdf?id=HJz9K7kJcX`, 2023.

[4] Surya Majumder and Adrija Bhattacharya. An automatic thaat and raga identification system using cnn-based models. *Innovations in Systems and Software Engineering*, October 2023.

[5] P. Sarath Kumar. Raga detection cnn model. https://www.kaggle.com/code/sarathkumarp/raga-detection-cnnmodel, 2022.

[6] Parampreet Singh and Vipul Arora. Explainable deep learning analysis for raga identification in indian art music, 2024. Accessed: 2025-05-02.