Curating a Public Carnatic Music Dataset: Scalable Extraction of Rāgam, Shruti, and Tālam Metadata for Computational Musicology

Sanjay Natesan*
sn2967@columbia.edu
Columbia University
New York, USA

Homayoon Beigi
hb87@columbia.edu
Columbia University
New York, USA

Abstract

We introduce a novel, publicly available corpus of South Indian *Carnatic* music, which—for the first time—spans **172** distinct *rāgams* (melodic frameworks) and **676** curated concert recordings, segmented into more than **11,219** audio clips. Each clip is annotated with its *shruti* (tonal center) and *tālam* (metrical cycle) as Linked-Data entities, enabling automatic interoperability with established Music Information Retrieval (MIR) ontologies [8]. The dataset was assembled through a hybrid pipeline that combines web-scale harvesting of YouTube concerts, automated signal processing for quality control, and expert-in-the-loop validation. To address inconsistencies in crowdsourced metadata, we introduce a pragmatic taxonomy that reconciles regional performance practices with canonical musicological literature [3, 9]. Case studies in automatic rāgam recognition and comparative tālam analysis illustrate how the resource advances computational musicology, cross-cultural MIR, and data quality assessment in digital libraries. This dataset is released under an open license on kaggle[1] and will be updated as the resource grows.

CCS Concepts

• **Applied computing** → **Sound and music computing**.

Keywords

Carnatic Music, Music Dataset, Rāgam Annotation, Shruti Detection, Tālam Classification, Music Information Retrieval, Digital Musicology, Cultural Heritage Preservation, Music Ontology

ACM Reference Format:
Sanjay Natesan and Homayoon Beigi. 2025. Curating a Public Carnatic Music Dataset: Scalable Extraction of Rāgam, Shruti, and Tālam Metadata for Computational Musicology. In *12th International Conference on Digital Libraries for Musicology (DLfM 2025), September 26, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3748336.3748354

*Now at Amazon, Seattle, USA
[1]https://www.kaggle.com/datasets/sanjaynatesan/carnatic-song-database

DLfM 2025, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2083-3/25/09
https://doi.org/10.1145/3748336.3748354

1 Introduction

The field of Music Information Retrieval (MIR) has seen a proliferation of large-scale, openly licensed corpora in recent years; however, most collections are biased towards Western music, resulting in a marked under-representation of non-Western traditions [8]. Recent efforts such as the multi-track Sanidha corpus for source separation [4] and the PIM-v1 archive for Hindustani music [10] signal growing interest in the computational analysis of South Asian music, yet no public Carnatic dataset to date spans the full breadth of melodic diversity documented by performers and pedagogues.

Building on our previously published TDNN-based rāgam identification study [6], we have substantially expanded both scope and metadata granularity. Our new collection includes **172** rāgams and **676** high-quality recordings, harvested from YouTube using a rigorous decision tree (Section 2), prioritizing audio fidelity, performance duration, and the avoidance of prolonged *thani avarthanam* (percussion solos). Well-defined selection relaxations ensured broad yet consistent coverage.

1.1 Why Digital-Library Infrastructure?

Non-Western music corpora often suffer from idiosyncratic metadata, hampering discoverability and reuse [8]. By aligning rāgam, shruti, and tālam labels under the guidance of CompMusic ontology framework, we propose a novel publishing that embraces FAIR principles and enables downstream tasks such as cross-collection discovery, Linked-Open-Data queries, and integration with score collections. Authority control and variation knowledge mitigates transliteration and regional spelling inconsistencies, a longstanding obstacle in South Indian music scholarship [3, 9].

1.2 Contributions

(1) **A Large-Scale Annotated Carnatic Corpus:** The first open dataset to encompass the *vast majority* of commonly performed rāgams, with shruti and tālam metadata linked to standard ontologies.
(2) **A Scalable Curation Workflow:** We detail a reproducible YouTube harvesting and expert-verification pipeline that balances automation with ethnomusicological rigor.
(3) **A Refined Rāgam Taxonomy:** We resolve inconsistencies in prior crowdsourced labels by proposing an ontology-compatible classification that accommodates regional variants without sacrificing canonical definitions.
(4) **Validation Through MIR Case Studies:** Experiments in automatic rāgam classification and rhythmic-cycle profiling

demonstrate the dataset's quality and its value for cross-cultural digital libraries.

2 Data Harvesting and Annotation Workflow

The dataset collection process followed a structured methodology to systematically harvest and annotate publicly available Carnatic music recordings from YouTube, ensuring comprehensive coverage and consistent quality. The workflow comprised automated scripting for harvesting, followed by meticulous manual validation.

2.1 Selection Criteria

We began by consulting Karnatik.com, an authoritative online database, to identify between one and three representative songs for each of the 172 rāgams, prioritizing popularity and canonical status [1]. Each identified song was then searched on YouTube, retrieving recordings adhering to strict quality standards:

- High audio fidelity with minimal distortion,
- Duration between 4 and 20 minutes,
- Absence of thani avarthanam,
- Balanced representation across both male and female shrutis.

In cases where no suitable recording was identified, we sequentially applied the relaxation criteria detailed below, in accordance with established methodologies for Carnatic dataset development [2]. This approach further safeguarded against disproportionate representation of any individual sample within the dataset.

(1) Select a recording of the chosen song with a duration between 3 and 4 minutes.
(2) If unavailable, select a recording with a duration between 20 and 40 minutes.
(3) Finally, consider a recording with a duration between 40 and 60 minutes.
(4) If still unavailable, select a defined quality recording performed by a vocalist in the range of the other sex.

2.2 Automated Harvesting

A dedicated Python script was developed to automate the downloading, segmentation, and validation of audio clips, utilizing YouTube's API alongside the `pytube`, `librosa`, and `soundfile` libraries. Metadata for each clip was meticulously logged in both structured spreadsheet and JSON formats, ensuring transparency and reproducibility throughout the workflow. To mitigate potential artifacts from song introductions or microphone issues, recordings were uniformly trimmed by removing the initial and final 10% of their duration. Subsequently, each file was segmented into 30-second excerpts, ultimately producing a corpus of more than 11,000 audio clips.

The workflow was designed to robustly manage exceptions, including unavailable videos and encoding anomalies, thereby maintaining resilience and continuity throughout the processing pipeline. Furthermore, its modular architecture facilitates adaptability to future changes in the dataset, such as the addition of new data points or the removal of existing videos due to them being taken offline.

2.3 Manual Annotation and Validation

Post-download, each clip underwent rigorous manual validation by Carnatic music experts. Annotations for shruti and tālam were verified and corrected as necessary, cross-referencing with authoritative musicological sources [3, 9]. Rāgam labels were checked for consistency, and spelling variations were scraped from the web [1].

This expert-in-the-loop methodology ensured high-fidelity audio and metadata, suitable for rigorous computational musicology research.

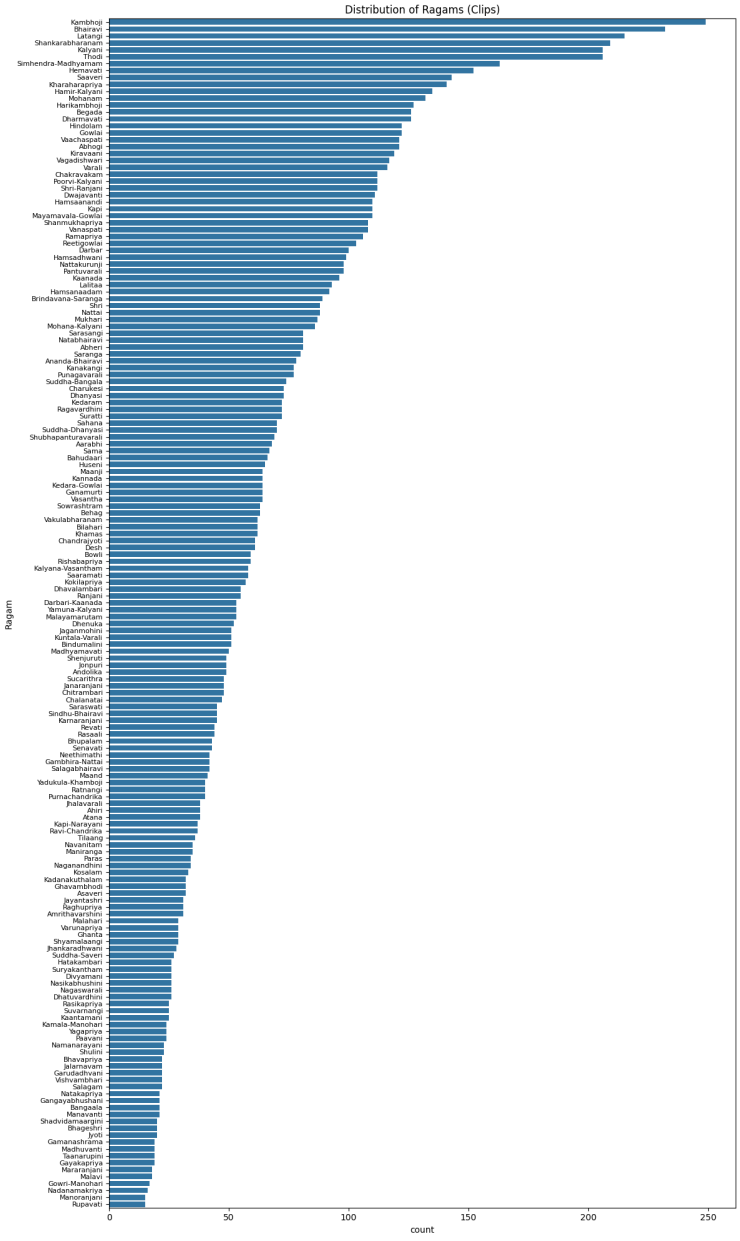

Figure 1: Distribution of Rāgams by Number of Clips

3 Dataset Description and Statistics

We provide a comprehensive statistical summary of the dataset, detailing the distribution of audio clips across various features.

Additionally, we present an outline of the JSON metadata structures, illustrating the interconnected ontology underpinning the dataset's relational organization.

3.1 JSON Metadata Structures

Three structured JSON schemas underpin the dataset metadata:

- **Song Metadata:** Each entry represents an individual song-clip with the following key fields:
  - `Song_ID`: Unique song identifier.
  - `Song_Name`: Title of the composition.
  - `Ragam`: Rāgam label (maps to `Ragam` in Rāgam Metadata).
  - `Composer`: Composer name.
  - `Singer_YN`: Indicates if a vocal rendition (Y/N).
  - `Accompaniment_YN`: Presence of non-tanpura accompaniment (Y/N).
  - `Shruthi`: Reference pitch.
  - `Talam`: Tālam label (maps to `Talam_Name` in Tālam Metadata).
  - `Youtube_Link`: Source link.
  - `Original_Song_Length`: Duration of song from Youtube in seconds.
  - `New_Song_Length`: Duration of song in seconds after preprocessing and trimming.
  - `Number_of_Clips`: Number of segmented 30-second clips.
- **Ragam Metadata:** Each entry describes a unique rāgam with fields such as:
  - `Ragam_ID`: Unique rāgam identifier.
  - `Ragam`: Canonical rāgam name.
  - `Melakarta`: Associated melakarta number.
  - `Is_Melakarta`: Indicator if the rāgam is a melakarta rāgam (Y/N).
  - `Variations`: Alternate spellings/transliterations.
  - `Karnatik_URI`: Reference URI from Karnatik.com[1].
  - `Arohanam`: Note sequences for ascending scales, sourced from Karnatik.com[1].
  - `Avarohanam`: Note sequence for descending scales, sourced from Karnatik.com[1].
- **Talam Metadata:** Each entry defines a rhythmic cycle with:
  - `Talam_ID`: Unique tālam identifier.
  - `Talam_Name`: Three-level identifier name.
  - `Unstructured_Talam_Name`: Canonical name (e.g., Adi, Rupakam).
  - `Kalai`: Value (usually 1 or 2).
  - `Total_Aksharam`: Number of aksharams (beats) per cycle.
  - `Edam_Offset`: Beat offset for entry.
  - `Variations`: Alternate spellings/transliterations.

These schemas ensure each song, rāgam, and tālam are uniquely identified, cross-referenced, and embedded with all metadata essential for robust computational analysis.

3.2 Rāgam Distribution

The distribution of clips per rāgam (see Figure 1) demonstrates the strong representation of rāgams such as Kambhoji, Bhairavi, and Shankarabharanam, which are frequently selected as main pieces in concerts and thus often performed in elaborate, multi-section presentations.

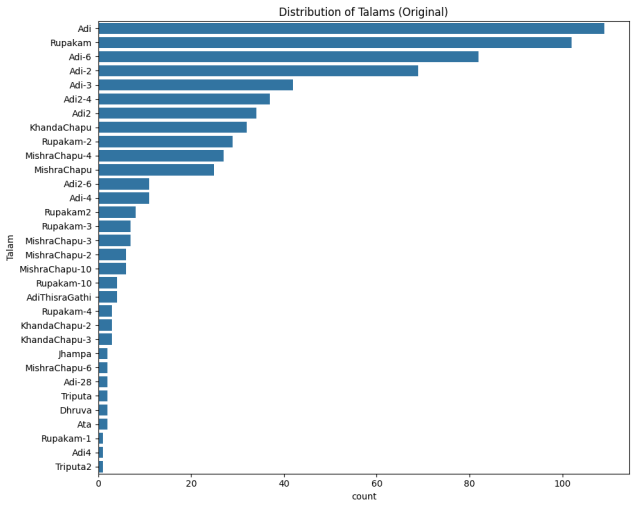

Figure 2: Distribution of Tālams (Full Songs, 3 Level Naming System)

3.3 Tālam Categorization and Distribution

Tālam annotations were described utilizing a three-level system:

- **Tālam Name:** e.g., Adi, Rupakam, Mishra Chapu, Khanda Chapu, Ata, etc.
- **Kālai:** Specifies the temporal resolution at which a tālam is performed. In one-*kālai*, each beat maintains its standard length, whereas in two-*kālai*, the duration of each beat is doubled, effectively slowing the tempo by half.
- **Aksharam Offset:** Number of *aksharams* (rhythmic subdivisions) offset in each performance, using pakkavadhyam (percussion) aksharam counting methodology [3].

Tālam categorization is detailed at multiple granularities. Figure 2 presents a distribution of full songs classified using all levels of the above-mentioned system. Figure 3 provides a count of clips based on only the tālam name. Both highlight the predominance of tālams such as Adi and Rupakam.

3.4 Melakarta Classification

Carnatic music organizes rāgams into two main categories: *melakarta* (parent) and *janya* (child) rāgams. Melakarta rāgams are the 72 foundational scales, each comprising seven ordered notes in both ascending (*arohana*) and descending (*avarohana*) sequences, and serve as the basis for all others in the tradition[3].

Janya rāgams are derived from melakartas, often by omitting, reordering, or varying notes, resulting in rich melodic diversity. They constitute most of the Carnatic repertoire featured in this dataset.

Within this founding version of the dataset, all 72 melakartas are represented, along with 100 janya rāgams. Figure 4 illustrates how melakartas like 22 and 28 yield numerous child rāgams, leading to their prominence in repertoire and in the dataset.

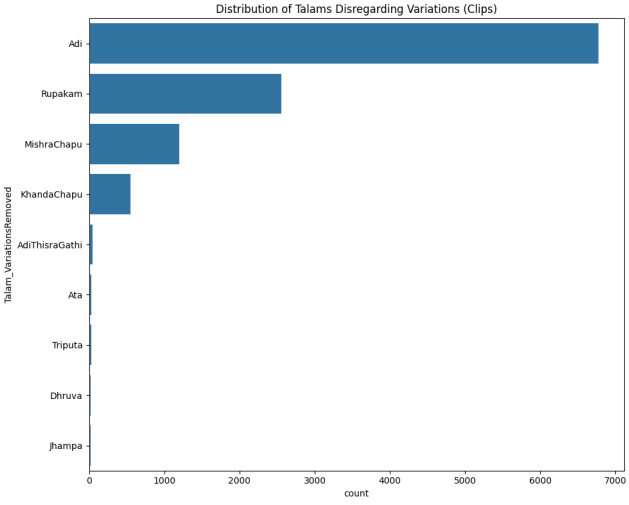

Figure 3: Distribution of Tālams (Song Clips, Single Level Naming System)

4 Computational Experiments and Case Studies

4.1 Baseline Model Evaluation

We evaluated our expanded and modified Carnatic dataset on both a hybrid LSTM/TDNN and an attention-based architecture. Both models were optimized for rāgam classification, and leverage a custom frequency binning approach [6].

Table 1 summarizes classification accuracy scores and benchmarks them against prior work:

Table 1: Baseline Model Performance Comparison

Reference	Dataset	Accuracy
Gulati et al. (2016) (LR) [2]	GCD	70.10%
Pillai & Mahajan (2017) (SVM) [7]	Melakartas	80.56%
Madhusudhan (2024) (LSTM-RNN) [5]	GCD	88.10%
Natesan (2024) (LSTM-TDNN) [6]	NCD	**95.31%**
Natesan (2024) (Attention) [6]	NCD	**99.27%**

These results demonstrate substantial improvement over previous methods and affirm the utility of the dataset for computational rāgam classification, mirroring outcomes in recent MIR literature [4][6].

5 Conclusion

We have presented a novel, open-access Carnatic music dataset, greatly expanding the scale and quality of digital resources for non-Western music analysis. By combining automated harvesting with detailed expert validation, we offer a resource that is both broad in scope and precise in metadata. Experimental evaluation with advanced neural architectures, yielded state-of-the-art results in rāgam classification.

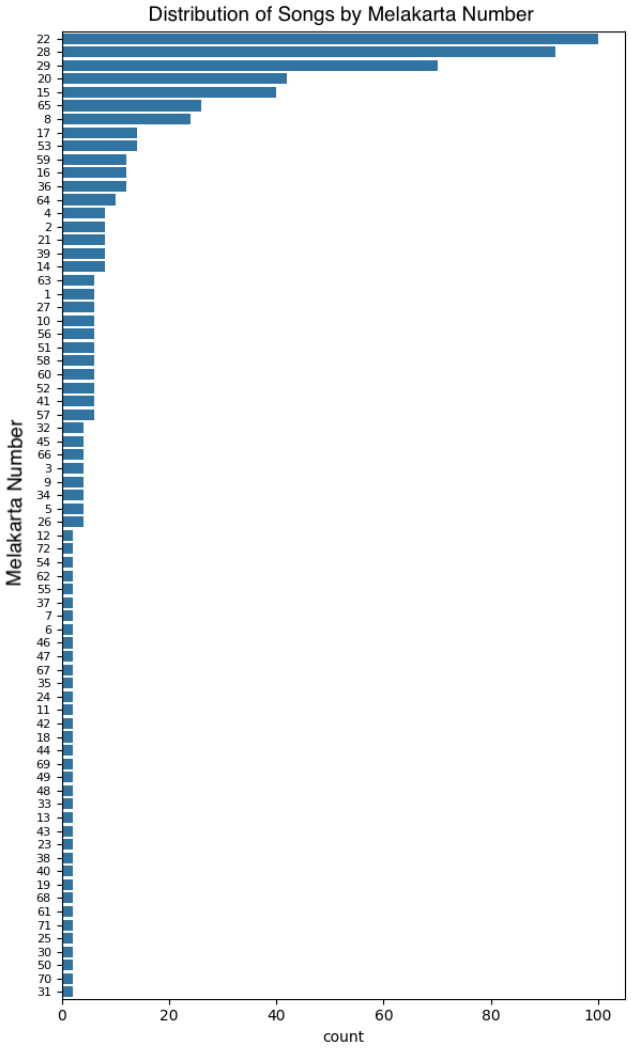

Figure 4: Distribution of Full Songs by Melakarta Number

Our dataset is intended as a foundation for future MIR research, digital musicology, and cross-cultural computational studies. Ongoing work includes expanding the corpus, refining metadata, and developing new experiments targeting rhythmic and tonal features unique to Carnatic music. By promoting open science and interoperability with MIR ontologies, we aim to foster broad engagement and scholarly innovation in computational ethnomusicology [6][8].

Acknowledgments

We thank colleagues, mentors, and reviewers for their valuable feedback on the manuscript. Editorial clarity was enhanced with the help of AI-based language tools, used solely for proofreading and readability; all substantive arguments and choices were made by the authors.

References

[1] 2025. *karnATik: Carnatic Music Resource.* https://www.karnatik.com/

[2] S. Gulati, J. Serrà, V. Ishwar, S. Şentürk, and X. Serra. 2016. Phrase-based Rāga Recognition Using Vector Space Modeling. In *IEEE ICASSP*. 66–70. doi:10.1109/ICASSP.2016.7471638

[3] T. M. Krishna and V. Ishwar. 2012. *Carnatic music: Svara, gamaka, motif and raga identity.* The Music Academy of Madras. http://hdl.handle.net/10230/20494

[4] Venkatakrishnan Vaidyanathapuram Krishnan, Noel Alben, Anish Nair, and Nathaniel Condit-Schultz. 2025. Sanidha: A Studio Quality Multi-Modal Dataset for Carnatic Music. arXiv:2501.06959 [cs.SD] https://arxiv.org/abs/2501.06959

[5] S. T. Madhusudhan and G. Chowdhary. 2024. DeepSRGM – Sequence Classification and Ranking in Indian Classical Music with Deep Learning. *Preprint* (2024). doi:10.48550/arXiv.2402.10168

[6] Sanjay Natesan and Homayoon Beigi. 2024. *Carnatic Raga Identification System Using Rigorous Time-Delay Neural Network.* Technical Report. Recognition Technologies, Inc. doi:10.13140/RG.2.2.17517.40164

[7] Rohan T. Pillai and Shrinivas P. Mahajan. 2017. Automatic carnatic raga identification using octave mapping and note quantization. In *2017 International Conference on Communication and Signal Processing (ICCSP)*. 0645–0649. doi:10.1109/ICCSP.2017.8286438

[8] Alastair Porter, Mohamed Sordo, and Xavier Serra. 2013. Dunya: A system for browsing audio music collections exploiting cultural context. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*. https://archives.ismir.net/ismir2013/paper/000179.pdf

[9] Subba V Rao. 1980. *Raganidhi: A comparative study of Hindustani and Karnatak Ragas Volume 1.* Music Academy.

[10] Parampreet Singh and Vipul Arora. 2025. Explainable Deep Learning Analysis for Raga Identification in Indian Art Music. *IEEE Transactions on Audio, Speech and Language Processing* 33 (2025), 2302–2311. doi:10.1109/taslpro.2025.3574839

Received 19 May 2025; accepted 26 June 2025