TABLA GHARĀNĀ RECOGNITION FROM AUDIO MUSIC RECORDINGS OF TABLA SOLO PERFORMANCES

Gowriprasad R[1] **Venkatesh V**[2] **Hema A Murthy**[1] **R Aravind**[1] **Sri Rama Murty K**[2]
[1] Indian Institute of Technology, Madras, [2] Indian Institute of Technology, Hyderabad, India
ee19d702@smail.iitm.ac.in, ee19mtech01010@iith.ac.in, hema@cse.iitm.ac.in
aravind@ee.iitm.ac.in, ksrm@ee.iith.ac.in

ABSTRACT

Tabla is a percussion instrument in Hindustani music tradition. Tabla learning and performance in the Indian subcontinent is based on stylistic schools called *gharānā- s*. Each gharānā is characterized by its unique style of playing technique, dynamics of tabla strokes, repertoire, compositions, and improvisations. Identifying the gharānā from a tabla performance is hence helpful to characterize the performance. This paper addresses the task of automatic gharānā recognition from solo tabla recordings. We motivate the problem and present different facets and challenges in the task. We present a comprehensive and diverse collection of over 16 hours of tabla solo recordings for the task. We propose an approach using deep learning models that use a combination of convolutional neural networks (CNN) and long short-term memory (LSTM) networks. The CNNs are used to extract gharānā discriminative features from the raw audio data. The LSTM networks are trained to classify the gharānā-s by processing the sequence of extracted features from CNNs. Our experiments on gharānā recognition include different lengths of audio data and comparison between various aspects of the task. An evaluation demonstrates promising results with the highest recognition accuracy of 93%.

1. INTRODUCTION

With the vast availability of varied music collections on the digital platform and widespread use of personal digital devices, there is a growing interest in accessing music based on its various characteristics. The limited availability of editorial metadata and annotations led to the need for music information retrieval to automatically extract music's characteristic properties from the audio recordings. Automatic identification of metadata from audio-like, stylistic school recognition—especially in the context of the same genre—is a tough task, even for humans.

This paper addresses the automatic identification of tabla gharānā-s, valuable metadata from solo tabla recordings. The percussion instrument *tabla* is an integral part of Hindustani music as it keeps track of rhythm. It is not only used as an accompaniment but also used in solo performances. Tabla solo is intricate and elaborate, with a variety of precomposed forms used for developing further elaborations based on the player's stylistic schools called *gharānā-s*. Identifying gharānā from a tabla performance provides valuable editorial metadata, which helps characterize performance and further musicological analysis. From the standpoint of Music Information Retrieval (MIR), studying and analyzing various tabla performance patterns is vital. It has applications in music description, auto-tagging, similarity measures, discovery, informed and enhanced music listening, music training, and computational musicology. We first discuss an overview of tabla and its gharānā-s.

1.1 Tabla and its gharānā-s

Tabla consists of a pair of drums, *dāyān* and *bāyān*, a treble drum and a bass drum respectively [1]. The tabla repertoire and technique are transmitted from generation to generation by guru-shishya (teacher-student) lineage [2], which is primarily an oral tradition. This guru-shishya lineage gave rise to different schools of tabla practice called gharānā-s. The word gharānā literally means the house of the teacher. Tabla solo performance showcases the percussionist's skill with tabla developing upon a variety of precomposed compositions such as *thēkā, kāyadā, palatā, rēlā, pēśkār and gaṭ* within the rhythmic framework called *tāl*. Each composition has different functional and aesthetic roles in a solo performance. The developments are intricate and elaborate based on the player's gharānā.

There are two major playing styles (*bāj*) in tabla, *bandh (closed) bāj* and *khulā (open) bāj* [3]. In *bandh bāj*, the tabla is played more on its border area; hence the stroke resonance is controlled or subdued. In this closed style of playing, importance is given to the sound of tabla and speedy progressions. In *khulā bāj*, the tabla is played more in the middle portion with open strokes using full palm and fingers. In this style, importance is given to tonal richness of the strokes with resonance. A few open and closed strokes are illustrated in the Figure 1. Based on these two broad playing styles, there are six *gharānā-s* developed namely *Delhi*, *Ajrada*, *Lucknow*, *Banaras*, *Farukhabad*, and *Punjab*. Each gharānā is characterized by its unique styles of playing, strokes played on the tabla, improvisations, and precomposed patterns. Every gharānā has a different approach to technique, and repertoire [3].

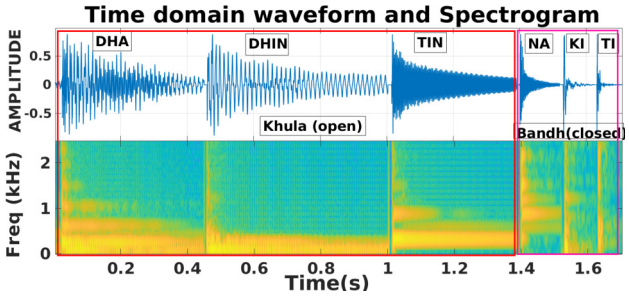

Figure 1. Illustration of a few tabla strokes.

1.2 Related Work

Most of the research related to Indian percussion focused mainly on stroke transcription and sequence modeling. Gillet *et al.* [4] focused on stroke transcription, and Chordia [1] extended the work by using additional features with larger-diverse dataset. Chordia et al. [5] used predictive models for tabla stroke sequence by making use of tabla syllables. Samudravijaya et al. [6] used hidden Markov models for recognition of tabla strokes. Kuriakose et al. [7] and Anantapadmanabhan et al. [8] worked on transcription of mridangam strokes. Chordia et al. [9] have worked on multiple viewpoints on modeling tabla sequences. Gupta et al. [10] identifying syllabic percussion patterns from the transcribed tabla audio.

The works on music style and classification are as follows. Vidwans et al. [11] used melodic contours to classify vocal style and classify cultural music using melodic features [12]. Agarwal et al. [13] did a comparative study of Indian and western music forms. Tang et al. [14] used hierarchical LSTMs for music genre classification. Gessle et al. [15] did a comparative analysis of Convolutional neural network (CNN) and Long short-term memory (LSTM) networks for music genre classification. Gogineni et al. [16] worked towards mridangam artiste identification from solo mridangam audio. We address a similar task of tabla gharānā recognition from solo tabla audios using the CNN-LSTM approach. As per our knowledge, identifying the tabla gharānā has not been attempted before.

The rest of the paper is organized as follows. The challenging factors influencing the task are categorically mentioned motivating the experiment. The dataset collected for the task is described. We formulate the task and explain the proposed model architecture. Multiple experiments addressing various facets of the task are described. The experimental results are analyzed and discussed.

2. TABLA GHARĀNĀ RECOGNITION

Tabla gharānā recognition is a task of identifying the stylistic schools of tabla given a solo tabla recording. We address the task by processing the composition-specific sequential information from the audio. We discuss the various aspects of the task and the collected dataset.

2.1 Motivation and Challenges

We motivate the task of tabla gharānā recognition by getting insight into the factors influencing the task. The factors include both the supporting and challenging aspects. Gottlieb [17] mentions three factors for comparing the similarities and differences in the playing: (1) Sound production, that is, quality and the technique used, (2) Repertoires, and (3) Rhythmic practices. The technique and the rhythmic practices differ from artiste to artiste. Compositions bearing the gharānā distinctions are based on the repertoires from each gharānā-s.

To get the expert's advice on the factors influencing the task, we consulted four tabla maestros. A few common opinions provided by the artistes on the task are mentioned here. Artistes nowadays would have learned from several teachers from different gharānā-s. Different gharānā styles will also influence their playing style [2]. Hence it is not straightforward to classify the artiste as coming from a particular gharānā in the present era. Therefore the repertoires [3] and compositions with some specific combination of certain strokes become the distinguishing factor in identifying the gharānā. This means that it is more straightforward and valid to recognize the gharānā based on the compositions rather than the artistes' playing styles.

Considering the vast diversity of the tabla solo repertoire and its practices, we list out a few possible challenging aspects influencing the system. (1) Tonic variability - Tabla is a pitched harmonic percussive instrument tuned to a specific tonic in a concert [18, 19]. As the tonic varies, the properties of the sound like harmonics, timbre, tone, etc., also vary. Thus the feature vectors representing the same stroke with different tonic will also change, challenging the system performance. In an ideal scenario, the gharānā distinctions are independent of tonic variability. (2) Artiste variability - Each artiste has their approach to techniques, finger postures, individual nuances with extempore development. Thus there exist an invariable influence of the artiste variability. (3) Composition variability - A few gharānā-s share similarities between their compositions. This is because some gharānā-s are the offshoots of others [2], share common stroke sets and techniques. There are equally significant instances where the compositional theme is entirely different within the gharānā itself. Thus composition variability has a strong influence on gharānā distinction.

2.2 Dataset Description

To experiment with various shades of tabla gharānā requires a diverse collection of annotated audio data. As there is no dataset available for this task, we collected solo tabla recordings from commercial audio CDs, live recordings from the artistes' archives, and online sources. This corpus consists of tabla solo from 18 different artistes with at least 20 years of tabla playing experience. The full-length tabla solo consists of different compositions played one after the another with a few cycles of ṭhēkās in between adjacent compositions, marking the start and end of a composition [3]. These compositions are usually from different gharānā-s. Thus we consider the aspect of repertoire, in so far as it has a bearing on the gharānā distinctions [2]. Kāyadā-s are the extendable compositions elaborated upon a theme from a particular gharānā through paltā-s. Paltā-s are the variations of original phrase or theme [2]. Each

Gharānā name (ID)	No. of Artistes	No. of Tonics	No.of Compositions	Durations hh:mm:ss
Ajrada (**A**)	7	5	21	2:23:58
Banaras (**B**)	8	9	35	3:09:01
Delhi (**D**)	10	5	16	2:32:11
Farrukhabad (**F**)	8	9	23	2:36:50
Lucknow (**L**)	7	7	36	2:34:52
Punjab (**P**)	7	5	42	2:53:27

Table 1. Dataset Description.

kāyadā rendition usually lasts for three to five minutes. The kāyadā compositions exists in all the gharānā-s [2]. Gaṭ-s and chakradhār are preset compositions that last for more than one to two metrical cycles. Hence in these compositions, one can observe the theme of the gharānā as well as the artiste's playing style.

Professional performers were employed to listen and extract the kāyadā, gaṭ, and chakradhār sections from the audio by marking the start and end points. In addition, the datasets from Gupta et al. [10], and Rohit et al. [20] are included in our dataset. Then four tabla maestros from different gharānā-s were requested to listen to these audio segments and give the ground truth gharānā labels. By doing so, we were able to collect around 16 hours of gharānā annotated audio. The details of the dataset are described in Table 1. The complete dataset is heterogeneous with artiste variability (18 artistes), tonic variability, tempo variability, and soft harmonium or sarangi accompaniment. Most of the audio played predominantly in tīntāl consisting of 16 beats. There are a few audios from the other tāls, such as jhaptāl and ektāl. The audios are sampled at 44.1 kHz. We use the downsampled version of the data at 16 kHz.

3. PROPOSED APPROACH FOR GHARĀNĀ RECOGNITION

The factors influencing the task necessitates a system that addresses various aspects and automatically perform gharānā recognition. With recent advances in Deep Neural Networks (DNNs), several tasks like music style recognition [21], music genre recognition [22] have achieved performance improvement with DNNs over the statistical methods. The success of DNNs motivates us to formulate a neural network based gharānā recognition system addressing various facets of the task.

We formulate the task by proposing a CNN-LSTM model. The model is trained by examples of different compositions from various artistes. The CNNs are trained to extract the local discriminative features pertaining to different stroke sounds from the raw audio. Tabla solos are developed, improvised, and elaborated upon a theme through a series of variations according to rhythmic practices [3]. Thus some strokes co-occur more often than others in the tabla compositions. Hence it is essential to train the models by encoding the sequence information. The LSTM networks are trained to classify the gharānā-s by processing the sequence of extracted features from CNNs.

3.1 Proposed model Architecture

The proposed CNN-LSTM method has a single training stage, which takes the raw samples and models posterior

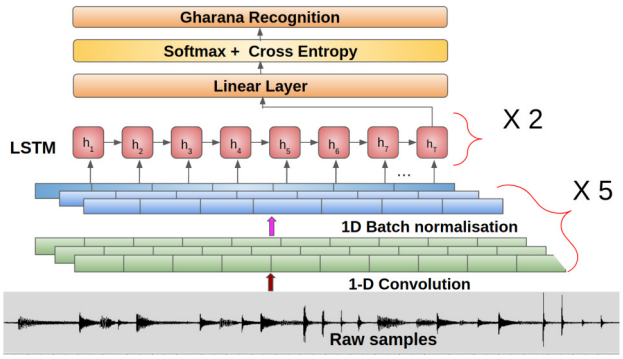

Figure 2. Proposed model architecture.

distribution over the six classes. The overall architecture of CNN-LSTM for gharānā recognition is shown in Figure 2. It has two components: a 1-dimensional convolution neural network (1D-CNN) and a long short-term memory (LSTM) network. Five layers of 1D-CNN are used in the model. The first 1D-CNN component takes the raw audio samples as input. The CNN section acts as a feature extractor, performing convolutions on samples along the time axis. It produces a feature vector for every 10 ms with a reduced frame rate. This sequence of feature vectors of 10 ms stride acts as input for the LSTM.

Every CNN layer is followed by BatchNormalization [23] and ReLU activation [24]. The kernel size, number of kernels, stride, and padding of each convolution layer are (1 x 10, 256, 5, 3), (256 x 8, 256, 4, 2), (256 x 4, 256, 2, 1), (256 x 4, 256, 2, 1), (256 x 4, 256, 2, 1), respectively. This 1D-CNN component configuration is adapted from contrastive predictive coding encoder [25]. The LSTM component consists of two-layer LSTMs, each with 256 dimensions. The final hidden activation of 256 dimensions is fed to the log-softmax classification layer with six units to get the gharānā predictions. The model is trained using Adam optimizer [26] with a batch size of 32 and the learning rate of 0.01. We have used learning schedule of dropping the learning rate to half of current whenever the validation loss doesn't decrease [27]. We also experimented by adding one more CNN with 5% dropout, and an LSTM layer. An implementation of the proposed architecture is available [1] .

4. EXPERIMENTS AND RESULTS

4.1 Human Assessment

For the task of gharānā recognition from audio recordings, we need to know the duration that encapsulates good gharānā specific information. From the experts feedback, we got to know that each gharānā has certain compositions practiced and played by almost all the tabla experts nowadays. Hence, if the tabla artistes were to recognize the gharānā by listening, they do it within two to three seconds if the composition is known. If they do not know the composition that is played, it takes some time to analyze the combination of strokes in the theme for a few cycles and predict the stylistic school gharānā. Thus to have a human assessment on time taken to recognize and the accuracy, a survey experiment is conducted.

[1] https://git.io/JEoo0

Student ID (Years of practice)	Avg Time taken (secs)	Accuracy
S1 (12)	9 - 12	85%
S2 (8)	11 - 15	76%
S3 (12)	7 - 10	87%
S4 (16)	6 - 8	97%
S5 (12)	8 - 10	92%
S6 (11)	8 - 11	89%
S7 (10)	10 - 13	83%
Average	8.4 - 11.3	87%

Table 2. Senior tabla students survey (Human baseline).

Since the four tabla maestros were also consulted for preparing the ground truth annotations, it is not proper to have the same experts for the human baseline survey experiment. Thus, we conducted a survey experiment on seven senior tabla students. These students have more than eight years of tabla practice. The students are asked to listen to the compositions from our collected dataset and predict the gharānā labels. The time taken for prediction is noted by pausing the audio at the instance of prediction. The details of the survey are tabulated in Table 2. One can observe that around 87% of the recordings were predicted correctly by the students on average. The average time taken to predict each recording was found to be eight to eleven seconds.

4.2 Baseline Experiments

As there are no previous works available on gharānā recognition, we adapted CNN maxpooling LSTM model from [22] to achieve baseline results. The model use two CNN layers and one LSTM layer. MFCC + derivatives + double derivatives (MFCC_$\Delta\Delta$) features are fed at the input. Each audio file is segmented to the 10-sec duration and treated as individual examples. MFCC_$\Delta\Delta$ features are extracted with a window size of 25 ms and a frameshift of 5 ms. Each example is represented by (2000, d) feature matrix. Each feature vector is of dimension d=57. It is the sum of MFCC (19)+ Δ+ $\Delta\Delta$ (d=19+19+19=57). 70% of the entire audio is used for training the models. 15% of the audio was used for development and test each. This model is considered as baseline for the following reasons: (1) The model is used for a similar task of music style classification, and a comparison with the state-of-the-art systems are made [22]. (2) This uses a simpler architecture similar to our proposed system. (3) Handcrafted MFCC_$\Delta\Delta$ features are used as the input.

4.3 CNN-LSTM Experiments

We conducted seven different sets of experiments with the proposed CNN-LSTM model addressing various aspects. The training, development, and test data split is the same as the baseline experiment $(70-15-15)$. The motivation for different experiments and their results are described.

4.3.1 Experiment with segment duration

Each theme is played initially at the speed of or double the speed of the original tempo for one or two cycles and is then played at four times the original tempo [3]. The underlying tempo is flexible in Indian music and differs from artiste to artiste. Some themes have a structure of three

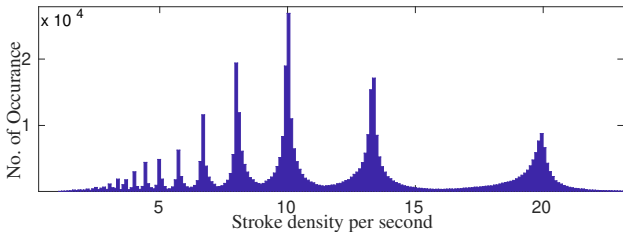

Figure 3. Histogram of stroke density per second, Mean=11.72, Median=10.25.

Weighted F1 Score						
Trial	**1**	**2**	**3**	**4**	**5**	**Avg**
3s	0.68	0.67	0.69	0.67	0.65	0.67
5s	0.79	0.77	0.78	0.8	0.79	0.78
10s	0.85	0.86	0.85	0.84	0.86	0.85
15s	0.85	0.90	0.89	0.88	0.91	0.89
15s DA	0.89	0.95	0.91	0.93	0.93	**0.92**

Table 3. Cross validation scores data length experiments.

strokes per beat (triple rhythm) [3, 18], which has nearly 12 strokes per beat at the fourth speed. Thus on average, around 8-12 strokes are played per beat at the fourth speed depending on the compositions. The mean and median of stroke density per second as obtained in the histogram, Figure 3 are found to be 11.72 and 10.25 respectively.

Repetition of the phrases is commonly observed in a musical concert. While playing at the fourth speed, one can hear at least one to two cycles of the theme or paltās in around 10 seconds of duration. This is evident as around 100 strokes are played in 10 seconds, which spans two to three cycles of the composition theme length. Since almost the whole structure of a theme can be found in about 10 seconds, it is adequate to segment the audio recordings into 10 to 15 second segment and treat them as separate examples. The average time duration as needed in the human assessment is also around 10 seconds. Thus we limited the length of train examples to be 10 to 15 seconds.

The entire data is segmented into 10s and 15s segments with 10% overlap with adjacent segments and treated as separate examples. The examples are randomly shuffled, and the train-dev-test split is done. Two models are trained and tested with 10s and 15s segments. The test segments are further chunked to 3s and 5s and treated as separate test examples in the next two experiments. This is done to analyze and get an insight into how much audio length is required for the system to recognize the gharānā specific sequence information. The longer the audio, the better the recognition. These experiments are 5-fold cross-validated, and the weighted F-1 scores are reported in Table 3. F-1 score is the harmonic mean of precision and recall.

4.3.2 Data Augmentation (DA)

To increase the training data diversity, data augmentation is performed. Speed perturbation with the factor of 0.9X and 1.1X (10% variation) without altering the pitch is done on the entire data using HPSS-TSM method [28, 29]. To get the value of the speed perturbation factor that does not alter the original data structure, we referred to the histogram of

Weighted F1 Score					
Exp	**Model~Variant**	**1**	**2**	**3**	**Avg**
15 IA	Std. Network	0.70	0.63	0.66	0.66
	+1 LSTM	**0.73**	0.68	0.70	0.70
15s IC	Std. Network	0.37	0.45	0.47	0.43
	+1 LSTM	0.38	0.46	**0.49**	0.44
15s IT	Std. Network	0.69	0.71	0.63	0.67
	+1 CNN	0.72	0.76	0.68	0.72
	f_0 = 260.63Hz	0.77	**0.80**	0.76	0.77

Table 4. Cross validation scores IA, IT, IC.

the average number of strokes per second.

Figure 3 shows the histogram depicting the statistical estimate of stroke density per second over the entire dataset. To get the stroke density per second, initially, the audio was preprocessed by computing the Hilbert envelop of linear prediction residual on the raw audio as described in [30]. Then the onset locations of each stroke are computed using a spectral flux onset detector [31]. The computed onset locations are considered for further analysis without any post-processing as the ground truth onset locations for the entire data are not available. The first difference of the onset locations gives the inter-onset interval (IOI). The inverse of IOI gives the stroke density.

The modes in the histogram have sharp peaks. One can infer that the underlying original tempo is almost uniform across the recordings. Thus a larger deviation from the original tempo is not advisable. The audio did not sound realistic if the speed perturbation was large. Therefore we restricted to only a 10% deviation from the original tempo. This increased the size of the dataset to three times the original. Since better performance was observed on 15s segments of the test data earlier, we choose to experiment with the same on augmented data as well.

4.3.3 Inter Artiste(IA) and Composition(IC) Experiments

The presentation of a particular composition by different artistes differs due to varied rhythmic practices and extempore development. We conducted IA experiment where the train and test artistes are distinct. Each compositional theme is unique in its own way. Thus to have an insight into various compositions, we performed an IC experiment. We took the help of senior students to have three sets where the compositions are diverse and performed the experiment. We experimented by adding a layer of LSTM to the model. This task is performed to check if an extra LSTM layer could learn more sequence information.

4.3.4 Inter Tonic (IT) Experiment

Different tonic essentially means the tabla itself is different. The tonic value is not a differentiating factor between the gharānā-s. We performed IT analysis to explore the diversity of the dataset. The train and test data tonics (tabla) are distinct in this experiment. We also tried two variants in IT experiments. (1) Adding one more CNN layer to the model with 5% dropout. This task is done to check if a bigger model could perform better in the IT experiment setup. (2) Preprocessing the data by normalizing the tonic of all the audio to 260.63 Hz (C_4), and then perform the experiment. The length of the audio is initially varied by a factor of $(\frac{260.63}{f_0})X$ using HPSS-TSM [28, 29] method, where f_0 is the original tonic value. Then it is resampled to play at the original speed, which in turn changes the pitch. This task is done to check if the preprocessing aids the performance in the IT experiment. The IA, IC, IT experiments are performed on 15s segments. The 3-fold cross-validation scores are reported in Table 4. The "Std.Network" in Table 4 refers to the proposed model with 5-CNN and 2-LSTM layers.

Method	**Models**	**Weighted F1**	**Accuracy**
Proposed CNN-LSTM	10s	0.85	0.86
	15s	0.89	**0.90**
	15s DA	**0.92**	**0.93**
	15s IA	0.66	0.67
	15s IC	0.43	0.45
	15s IT	0.67	0.67
	15s IA + DA	0.69	0.71
	15s IC + DA	0.44	0.45
	15s IT + DA	0.7	0.7
Baseline (Sec 4.2)	CNN-LSTM 10s-(MFCC)	0.53	0.55

Table 5. Results with different experiments.

5. ANALYSIS AND DISCUSSION

5.1 Performance analysis

Table 5 summarizes the experiments' average performance scores after n-fold cross-validation. One can observe that the proposed CNN-LSTM model tested with 15s data segments has the best performance of 90%. Performance dropped during IA, IC, and IT experiments. This is evident as the experiments are on the challenging aspects influencing the system. From Table 4, we can see that the addition of an extra CNN layer improved the performance of IT by 5%. An extra LSTM layer improved the performance of IA by 4% and IC by 1%. Data augmentation improved the performance of 15s, IA , IT by 3% and IC by 1%. This indicates that tuning a bigger model aids the system to be robust to tonic and artiste variation. This emphasizes that the composition variability has a larger influence on the task than the artiste variability. This is also evident from the artistes' feedback described in section 2.1.

The experiment with preprocessing the data by normalizing the tonic favored but did not add a larger value. If this preprocessing approach helped, the results should have been on par with the 15s experiment with the random train-test split. Few reasons which we are able to examine are as follows. (1) The artifacts and distortions due to time scale modification and resampling. (2) The bass drum is usually not set to any tonic. Shifting the pitch of the treble drum invariably shifts the pitch of the bass drum with the same factor. This does not sound realistic. It is to be observed that the performance improved nearly by 10% by preprocessing the audio. This indicates that suitable preprocessing aids the system performance. We observed that adding one more CNN and LSTM layer to the proposed model did not improve the performance in the random shuffle exper-

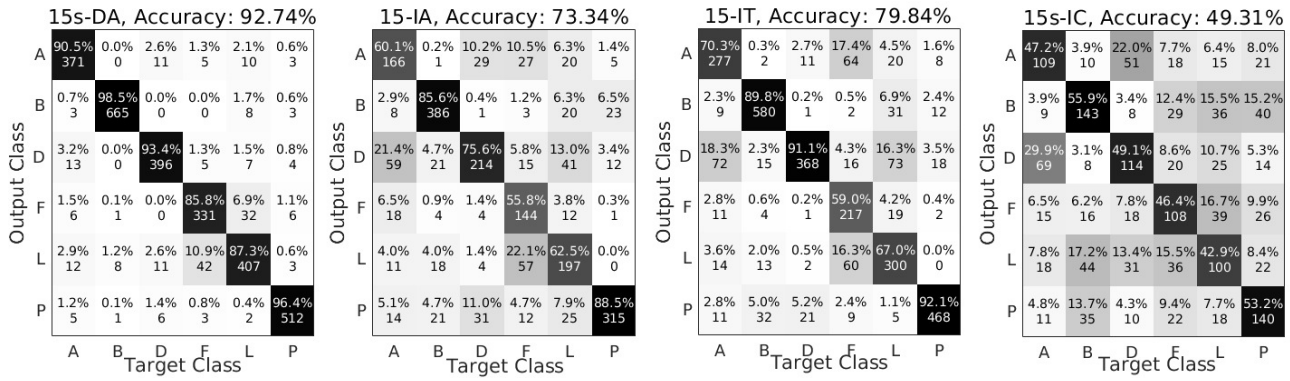

Figure 4. Confusion matrices depicting the best performing experiments on different facets.

iments with segment duration. Thus we reported the average scores from the proposed standard model in Table 5.

One can observe a clear improvement in the performance by 32% by comparing the 10s experiment with the baseline. This indicates that the features extracted from the raw audio by the CNN model have better discriminatory information than the handcrafted MFCC features.

5.2 Confusion analysis

The confusion matrices depicting the results for the major experiments are displayed in Figure 4. Figure 5 shows the t-SNE visualization of test data embedding from the CNN-LSTM model extracted from the 15s-DA samples. These embeddings are tapped at the output of the LSTM layer just before the linear layer. The visualization of embeddings and confusion matrices depicts the similarities and uniqueness of gharānā-s among each other. Different clusters are marked for the benefit of reference. One can observe that Punjab gharānā embeddings form a clear separation from the others (C1). This is evident as the Punjab gharānā has had a separate existence and is unique compared to the other ones [3].

We can observe an inevitable overlap and confusion between Lucknow and Farrukhabad embeddings (C2) as well as Delhi and Ajrada embeddings (C3). This overlap is acceptable as the Ajrada, and Farrukhabad gharānā-s are the offshoots of Delhi and Lucknow gharānā-s respectively [2]. It is also a fact that Delhi and Ajrada have similar playing styles, and both belong to bandh bāj [3]. Lucknow and Farrukhabad also lot of similarities traditionally, and both belong to khulā (open) bāj [17].

The founders of Lucknow gharānā hailed from Delhi gharānā [2]. Both Farrukhabad and Banaras styles originated from Lucknow gharānā. Thus we can observe the Lucknow gharānā embeddings getting overlapped with others. This confusion is also observed in various experiments addressing different facets. One can also observe the clear separation of Banaras gharānā embeddings (C4). This is evident as Banaras gharānā has a lot of changes to Lucknow gharānā getting influenced by Pakhawaj bōls [2, 17] and has a unique way of playing the bass drum [3]. Farrukhabad repertoire exploited entire vocabulary of the instrument [3]. Thus Farrukhabad embeddings can be seen confused with others.

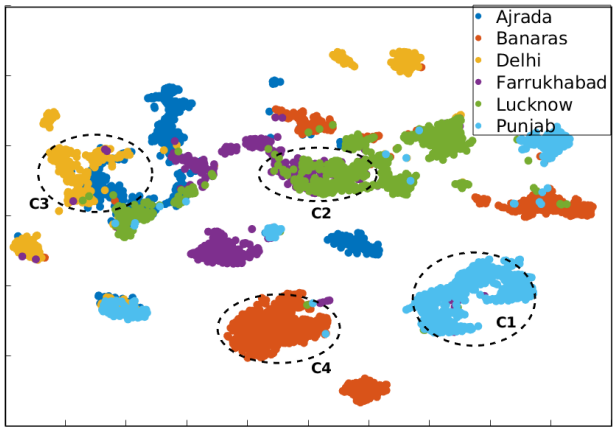

Figure 5. t-SNE visualization of test data embedding extracted from the 15s CNN-LSTM Model.

By visualizing the embeddings, and confusion matrices, one can verify many traditional similarities and differences between the gharānā-s as depicted and explained in the sources [2, 3, 17, 18]. Thus one can claim that the model has been trained in a positive way.

6. CONCLUSION

We addressed an unexplored problem of recognizing tabla gharānā, the stylistic schools of tabla, by proposing a deep learning model. The task used around 16 hours of gharānā class annotated data consisting of various compositions played by contemporary artistes. We motivate the problem and present different facets and challenges in the task. CNN and LSTMs in tandem are trained to extract gharānā discriminative features from the raw audio data and classify gharānā-s by processing the sequence information. Different experiments addressing various aspects of the task are performed. Additionally, proposed variants improved the performance in respective experiments. The system performance is comparable with the human assessment. The proposed CNN-LSTM model delivered promising results with the highest accuracy of 93% and a relative improvement of 31% over the considered baseline.

As a first attempt, we started by proposing a CNN-LSTM model for the task. Incorporating the tonic information during training will be beneficial. We aim to extend the dataset by incorporating other tabla compositions such as pēśkār, tukda, etc., and perform the experiments.

7. ACKNOWLEDGMENTS

The authors are grateful to the tabla maestros Ramesh Dhannur, Kiran Yavagal, Aneesh Pradhan, Aditya Kalyankur, and Anirudh Sharma for their support and help. We are thankful to Ajay Srinivasamurthy for his support and timely guidance. We thank Prof. Preeti Rao and Swapnil Gupta for sharing their dataset with us.

8. REFERENCES

[1] P. Chordia, "Segmentation and recognition of tabla strokes." in *Proc. 6th International Society for Music Information Retrieval (ISMIR), 2005*.

[2] S. Bagchee, *Nād: Understanding rāga music*. Eeshwar, 1998.

[3] A. Pradhan, *Tabla: A Performer's Perspective*. BookBaby, 2011.

[4] O. Gillet and Richard, "Automatic labelling of tabla signals," in *Proc. 4th International Society for Music Information Retrieval (ISMIR), 2003, Baltimore,USA*.

[5] P. Chordia, A. Sastry, and S. Şentürk, "Predictive tabla modelling using variable-length markov and hidden markov models," *Journal of New Music Research*, vol. 40, no. 2, pp. 105–118, 2011.

[6] K. Samudravijaya, S. Shah, and P. Pandya, "Computer recognition of tabla bols," Technical report, Tata Institute of Fundamental Research, Tech. Rep., 2004.

[7] J. Kuriakose, J. C. Kumar, P. Sarala, H. A. Murthy, and U. K. Sivaraman, "Akshara transcription of mrudangam strokes in carnatic music," in *Twenty First National Conference on Communications (NCC) 2015*.

[8] A. Anantapadmanabhan, A. Bellur, and H. A. Murthy, "Modal analysis and transcription of strokes of the mridangam using non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2013*.

[9] P. Chordia, A. Sastry, T. Mallikarjuna, and A. Albin, "Multiple viewpoints modeling of tabla sequences." in *Proc. 11th International Society for Music Information Retrieval (ISMIR), 2010*, p. 11th.

[10] S. Gupta, A. Srinivasamurthy, M. Kumar, H. A. Murthy, and X. Serra, "Discovery of syllabic percussion patterns in tabla solo recordings," in *Proc. 16th International Society for Music Information Retrieval (ISMIR); 2015 Oct 26-30; Málaga, Spain.[Málaga]. p. 385-391*.

[11] A. Vidwans, K. K. Ganguli, and P. Rao, "Classification of indian classical vocal styles from melodic contours," in *Serra X, Rao P, Murthy H, Bozkurt B, editors. Proceedings of the 2nd CompMusic Workshop; 2012 Jul 12-13; Istanbul, Turkey. Barcelona: Universitat Pompeu Fabra; 2012. p. 139-146*, 2012.

[12] A. Vidwans, P. Verma, and P. Rao, "Classifying cultural music using melodic features," in *2020 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2020, pp. 1–5.

[13] P. Agarwal, H. Karnick, and B. Raj, "A comparative study of indian and western music forms." in *Proc. 14th International Society for Music Information Retrieval (ISMIR)*, 2013, pp. 29–34.

[14] C. P. Tang, K. L. Chui, Y. K. Yu, Z. Zeng, and K. H. Wong, "Music genre classification using a hierarchical long short term memory (lstm) model," in *Third International Workshop on Pattern Recognition*, vol. 10828. International Society for Optics and Photonics, 2018, p. 108281B.

[15] G. Gessle and S. Åkesson, "A comparative analysis of cnn and lstm for music genre classification," 2019.

[16] K. Gogineni, J. Kuriakose, and H. A. Murthy, "Mridangam artist identification from taniavartanam audio," in *Twenty Fourth National Conference on Communications (NCC) 2018*. IEEE, pp. 1–6.

[17] R. S. Gottlieb, *Solo tabla drumming of North India: Its repertoire, styles, and performance practices*. Motilal Banarsidass Publishers, 1993.

[18] S. K. Saxena, *The Art of Tablā Rhythm: Essentials, Tradition, and Creativity*. Sangeet Natak Akademi, 2006, no. 8.

[19] A. Anantapadmanabhan, J. Bello, R. Krishnan, and H. Murthy, "Tonic-independent stroke transcription of the mridangam," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society, 2014.

[20] M. Rohit and P. Rao, "Acoustic-prosodic features of tabla bol recitation and correspondence with the tabla imitation." in *Interspeech*, 2018, pp. 1229–1233.

[21] B. Kumaraswamy and P. Poonacha, "Deep convolutional neural network for musical genre classification via new self adaptive sea lion optimization," *Applied Soft Computing*, p. 107446, 2021.

[22] D. Ghosal and M. H. Kolekar, "Music genre recognition using deep neural networks and transfer learning." in *Interspeech*, 2018, pp. 2087–2091.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[24] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Icml*, 2010.

[25] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] K. Audhkhasi, A. Rosenberg, A. Sethy, B. Ramabhadran, and B. Kingsbury, "End-to-end asr-free keyword search from speech," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1351–1359, 2017.

[28] J. Driedger, M. Müller, and S. Ewert, "Improving time-scale modification of music signals using harmonic-percussive separation," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 105–109, 2014.

[29] S. Yong, S. Choi, and J. Nam, "Pytsmod: A python implementation of time-scale modification algorithm," in *Extended Abstracts for the Late-Breaking Demo Session of the 21th International Society for Music InformationRetrieval Conference (ISMIR), 2020*.

[30] R. Gowriprasad and K. S. R. Murty, "Onset detection of tabla strokes using lp analysis," in *International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2020, pp. 1–5.

[31] S. Dixon, "Simple spectrum-based onset detection," *MIREX 2006*, p. 62, 2006.