International Journal of Applied Mathematics
Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (online version)

CROSS-CULTURAL MUSIC GENRE CLASSIFICATION WITH MACHINE AND DEEP LEARNING

Pushparajan. M[1], K.T. Sreekumar[2], K.I. Ramachandran[3*], C Santhosh Kumar[4]

[1]Assistant Professor, Department of Mechanical Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India, Email ID: m_pushparajan@cb.amrita.edu
[2]Assistant Professor, Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India, Email ID: kt_sreekumar@cb.amrita.edu,
Orcid ID: 0000-0003-3915-6128
[3]*Professor, Department of Mechanical Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India, Email ID: ki_ram@cb.amrita.edu,
Orcid ID: 0000-0001-5013-2886
[4]Professor, Department of Electronics & Communication Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, India,
Email ID: cs_kumar@cb.amrita.edu, Orcid ID: 0000-0002-7825-0179

Abstract

Music genre classification (MGC) is one of the most engaging challenges in the field of music information retrieval, especially when music comes from diverse cultural traditions. In this study, we bring together Western, Hindustani, and Carnatic musical forms to explore how both traditional machine learning models and modern deep learning architectures can understand and classify music styles. Using three datasets, GTZAN for Western genres and two Indian datasets that we compiled to represent Hindustani and Carnatic ragas, we experimented with audio segments of 30, 15, 10, 6, and 3 seconds to study how segment length affects classification performance. In the first part, we applied well-known machine learning algorithms such as logistic regression, random forest, gradient boosting, support vector machine (SVM), and artificial neural network (ANN). Among these, SVM and ANN performed best, and further enhancements using principal component analysis (PCA) and XGBoost improved their accuracy. In the second part, we focused on deep learning: a convolutional neural network (CNN) trained on mel-spectrograms, followed by a proposed convolutional neural network-long short-term memory (CNN-LSTM) hybrid model that captures both the spectral and temporal patterns of music. This model consistently outperformed all other approaches. We achieved classification accuracies of 89.63%, 99.09%, and 97.95% for Western, Hindustani, and Carnatic genres, respectively. The findings highlight how audio duration, feature choice, and temporal modeling together influence cross-cultural music classification.

Key Words and Phrases: Music genre classification, Music information retrieval, Convolutional neural networks, Long short-term memory, Mel-spectrogram.

1 Introduction

Among the many manifestations of human creativity, music stands out as deeply expressive, and categorizing it by genre continues to pose a significant challenge in music information retrieval (MIR) [8]. Music genre classification (MGC) serves as a fundamental component in modern music recommendation systems, facilitating personalized content delivery to listeners, organizing vast digital libraries [12], and even supporting musicological research. However, music is not universal in style or structure, what works well for Western pop or jazz may not hold for Indian classical traditions such as Hindustani or Carnatic music. The diversity of instruments, rhythms, tonal systems, and improvisational nuances across these cultures makes cross-cultural classification an

Received: August 07, 2025

Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (online version)

especially meaningful problem to explore.

In this study, we look beyond a single dataset or tradition to examine how both machine learning and deep learning techniques perform across different musical worlds. By bringing together three datasets, the well-known Western GTZAN corpus [16] and two Indian classical datasets (Hindustani and Carnatic) that we developed, we aim to understand how various algorithms learn to distinguish between contrasting styles and structures. Each dataset is segmented into clips of different lengths (30, 15, 10, 6, and 3 seconds) to study how context duration influences classification.

Despite substantial progress in the field, most published works focus only on Western datasets and often evaluate models under inconsistent conditions. Moreover, the role of audio segment duration, a crucial factor in real-world applications like short-clip tagging, is rarely discussed. Our work seeks to fill these gaps by offering a unified experimental framework that bridges classical and contemporary approaches while including culturally diverse datasets.

We introduce a cross-cultural MGC benchmark combining GTZAN with two curated Indian classical datasets, representing Hindustani and Carnatic traditions. A comparison of traditional machine learning (ML) classifiers (logistic regression (LR) [3], random forest (RF), gradient boosting (GB), support vector machine (SVM) [6], and artificial neural network (ANN)) and deep learning (DL) architectures (convolutional neural network (CNN) [13], and the proposed convolutional neural network-long short-term memory (CNN-LSTM) model) is presented under the same preprocessing and evaluation protocols. We analyze how feature representation, segment duration, and temporal modeling affect performance across different styles of music. Results are visualized through confusion matrices, highlighting where models succeed and where they struggle.

Section 2 reviews related research in MGC and MIR. Section 3 discusses datasets and preprocessing. Section 4 explains the machine and deep learning methods and the proposed method. Section 5 describes the experiments and results, followed by conclusion in Sections 6.

2 Related Work

In music information retrieval (MIR) [8] [14], one of the fundamental challenges is to categorize audio signals automatically according to their genres. Initial approaches mainly utilized manually engineered features, such as chromagrams and mel-frequency cepstral coefficients (MFCCs), along with conventional machine learning models, including support vector machines (SVM) and hidden Markov models (HMM) [2] [15]. A notable contribution by George Tzanetakis et al. [16] introduced a method for automatic genre classification using features like timbral texture, pitch content and rhythmic content features. Panagakis et al. [11] introduced a joint sparse low-rank representation framework that, by denoising test samples and uncovering latent subspaces, outperformed sparse representation, nearest-neighbor baselines, and nearest-subspace. Expanding the research focus to non-Western music, Sakinat O. Folorunso et al. [5] explored Nigerian MGC using models like random forest, SVM, extreme gradient boosting, and k-nearest neighbor. While these conventional machine learning approaches have shown reasonable effectiveness in genre classification, their performance often falls short when applied to diverse and complex music datasets, highlighting the need for more advanced methods.

Deep learning techniques have brought substantial improvements to the task of MGC. CNNs [10] have become the leading framework in MGC, demonstrating remarkable performance when trained on spectrogram representations of audio signals [4]. In their study, Wang et al. [7] proposed a deep

Received: August 07, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (online version)

learning driven approach for music genre analysis, achieving notable improvements in both accuracy and efficiency by using sophisticated feature extraction and classification techniques. Bahuleyan et al. [1] contributed to MGC field by comparing deep learning and traditional machine learning approaches. Liu et al. [9] developed a convolutional neural network framework for music genre classification that combines detailed low-level spectrogram features with mechanisms to model long-range temporal context, resulting in improved classification accuracy.

While CNNs effectively learn local time-frequency patterns, they struggle to model long-range dependencies, thereby limiting their capacity to represent the overall temporal and structural characteristics of music. To address this limitation, we propose a CNN-LSTM model, using the strengths of CNN and LSTM. This approach aims to achieve state-of-the-art performance in music genre classification while maintaining computational efficiency. By effectively capturing both the local and global dependencies within the music signal, our approach addresses the limitations of the existing methods.

3 Datasets and Preprocessing

Our study brings together three diverse and complementary datasets to explore genre and raga-family classification across Western and Indian musical traditions.

Western music dataset: GTZAN is a widely used benchmark for western genre classification, developed by G. Tzanetakis and P. Cook [16]. It includes 10 genres: blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, and rock. Each genre comprises 100 tracks, 30 seconds in length, preserved in uncompressed .wav format with 22.05 kHz sampling rate and 352 kbps bit rate.

Hindustani music dataset: To represent North Indian classical music, we compiled a Hindustani music dataset containing 10 ragas, each raga with a total of 50 minutes duration, taken from Pandit Bhimsen Joshi's renditions. The ragas are bhairav, bhimpalasi, darbari, durga, lalit, malkauns, marwa, tilakkamod, todi, & yaman.

Carnatic music dataset: For South Indian classical music, we created a Carnatic music dataset comprising 20 ragas, each raga with a total of 50 minutes duration, taken from Dr.M. Balamuralikrishna's renditions. The ragas are abheri, abhogi, amritavarshini, anandabhairavi, arabhi, charukeshi, hamsanandi, hindolam, kalyani, kamboji, kharaharapriya, mayamalavagowla, mohanam, pantuvarali, revati, shankarabharanam, shanmukhapriya, simhendramadhyamam, subhapantuvarali, & thodi.

Every track from the datasets was divided into shorter audio segments of 30 s, 15 s, 10 s, 6 s, and 3 s. For the machine learning models, we extracted time domain features (tempogram, central moments, signal energy, root mean square energy, and zero crossing rate) and frequency domain features (MFCCs, chroma features, spectral centroid, spectral entropy, spectral roll-off, and spectral contrast). For the deep learning models, we computed mel-spectrograms using a 1024-point FFT with a 50% hop length and 128 mel bands, applying logarithmic scaling to emphasize perceptually meaningful features.

4 Methodology

4.1 Machine Learning Classifiers and Dimensionality Reduction

In this study, we employed a range of well-established machine learning classifiers to build strong

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (online version)

and diverse baselines. These included logistic regression (LR), random forest (RF), gradient boosting (GB), support vector machine (SVM), and artificial neural network (ANN). Each model was trained using the same set of extracted time and frequency domain features to ensure a fair comparison across methods.

To enhance the performance of these classifiers and reduce redundancy among features, we used principal component analysis (PCA) for dimensionality reduction. Additionally, XGBoost, a powerful gradient-boosted tree-based algorithm, was applied both as an independent classifier and as a feature selection tool to identify the most informative components. These dimensionality reduction and selection methods helped streamline computation, reduce overfitting, and emphasize features that carried the most discriminative power across genres and raga families.

4.2 Convolutional Neural Network (CNN)

Deep learning models like CNNs [17] excel due to their capacity to learn from raw input data. By utilizing convolutional layers, CNNs selectively concentrate on distinct input segments, rendering them particularly effective for image and audio data processing. These layers use learnable filters, represented as $W_f \in \mathrm{R}^{k \times k \times c_i}$, where $k \times k$ represents the filter size, and c_i denotes the number of input channels, respectively. The convolution operation between the filter and the input feature map $X^l \in \mathrm{R}^{M \times N \times c_i}$ is:

$$F_m^n(x,y) = \sum_{i=1}^{c_i} \sum_{p=-\frac{k-1}{2}}^{\frac{k-1}{2}} \sum_{q=-\frac{k-1}{2}}^{\frac{k-1}{2}} W_f^{n,i}(p,q) X^l_{x+p,y+q,i} + b_m^n \qquad (1)$$

where, F_m^n represents the feature map at layer l after convolution with the n-th filter, and b_m^n denotes the bias term. To effectively classify music genres, CNNs employ a hierarchical architecture. Convolutional layers initially extract local patterns from spectrograms, followed by pooling layers that reduce feature map size while retaining vital information, ultimately enabling the network to learn sophisticated feature representations. Figure 1 shows the architecture of the CNN system.

4.3 Proposed CNN-LSTM Network

The CNN effectively captures spatial patterns in spectrograms, but music also contains important temporal dynamics that unfold overtime. To address this, we proposed a hybrid CNN-LSTM architecture that combines convolutional and recurrent layers in a single framework. The CNN front end extracts compact spectral features, while the long short-term memory (LSTM) layer models how these features evolve across time, enabling the network to understand melodic and rhythmic progression. This design allows the model to learn both the texture and flow of music, and also how tones, harmonies, and rhythmic accents develop across a phrase or composition. Such an approach is particularly valuable for Indian classical music, where improvisation and gradual melodic development are key characteristics.

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (online version)

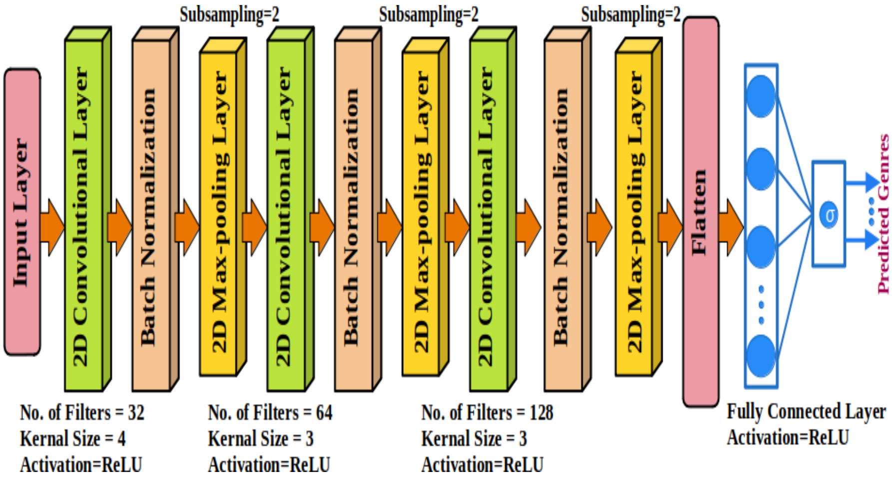

Figure 1: Architecture of CNN.

The convolutional operation in a CNN can be represented mathematically as follows:

$$\text{output}[i, j] = f\left(\sum_{m,n} \text{input}[i+m, j+n] \times \text{filter}[m, n]\right) \qquad (2)$$

where output[i, j] is the output at location (i, j), input[$i + m, j + n$] is the input at location ($i + m, j + n$), filter[m, n] is the filter, and f is an activation function (e.g., ReLU). Long short-term memory layers are designed to model temporal dependencies, employing gating mechanisms to control information retention and update across time steps:

forget gate = $\sigma(W_f \cdot [h_{\text{prev}}, x_t])$ (3)

input gate = $\sigma(W_i \cdot [h_{\text{prev}}, x_t])$ (4)

output gate = $\sigma(W_o \cdot [h_{\text{prev}}, x_t])$ (5)
cell state candidate = $\tanh(W_c \cdot [h_{\text{prev}}, x_t])$ (6)
cell state = forget gate·cell state prev+input gate·cell state candidate (7)
hidden state = output gate · tanh(cell state) (8)
where σ is the sigmoid activation function, tanh is the hyperbolic tangent function, W_f, W_i, W_o, and W_c are weight matrices, h_{prev} is the previous hidden state, x_t is the current input, and cell_state_prev is the previous cell state.

Finally, the fūlly cōnnected layers integrate spatial and temporal features into a fixed-length representation, while the output layer uses a softmax activation function to assign genre probabilities. This combination of convolutional and recurrent mechanisms enables CNN-LSTM models to robustly learn both local and global audio features, significantly improving genre classification accuracy. The architecture of the CNN-LSTM system is shown in Figure 2.

International Journal of Applied Mathematics
Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (online version)

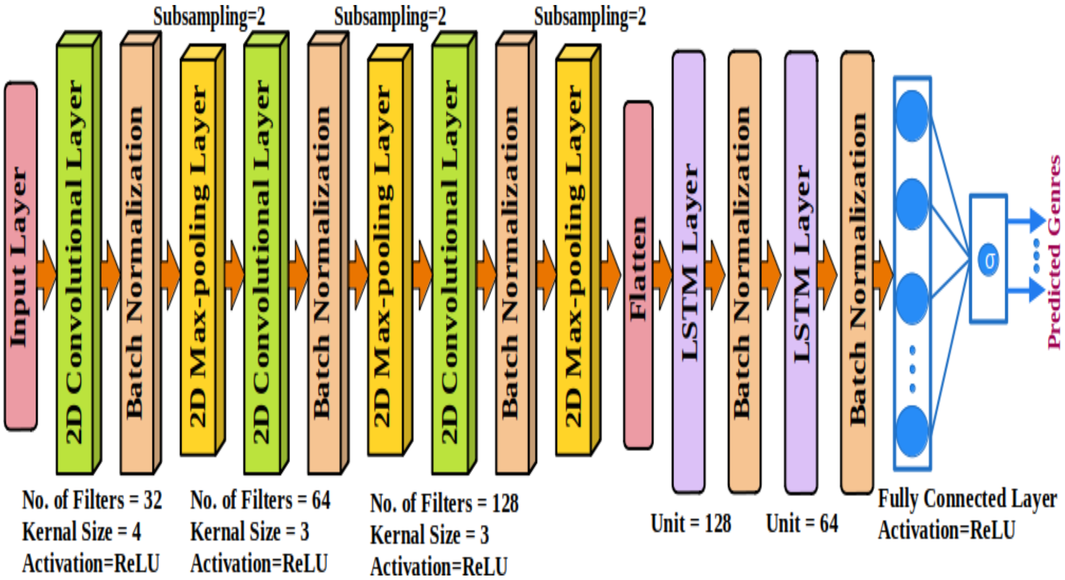

Figure 2: Architecture of CNN-LSTM.

5 Experiments and Results

All experiments were conducted under a consistent framework across the three datasets. The results from our experiments provide a clear picture of how both machine learning and deep learning approaches perform across Western and Indian musical traditions. Overall, the deep learning models, especially the proposed CNN-LSTM, demonstrated superior performance across all datasets and segment durations.

5.1 Machine learning based MGC systems

The performance accuracies of the baseline MGC systems are tabulated in Table 1. When comparing all models, SVM and ANN consistently stood out among the machine learning techniques, achieving strong results on both Western and Indian datasets. Applying the feature selection algorithms PCA and XGB improved their stability and accuracy, by reducing redundancy among features, for which the results are tabulated in Table 2.

Table 1: Performance evaluation of baseline MGC system.

Classifiers	Western					Hindustani					Carnatic				
	Accuracy (%)					Accuracy (%)					Accuracy (%)				
	30s	15s	10s	6s	3s	30s	15s	10s	6s	3s	30s	15s	10s	6s	3s
LR	76	74	74	73	71	94	93	92	89	87	74	72	73	73	71
RF	77	76	75	75	73	90	88	87	86	85	76	76	76	75	73
GB	76	76	76	75	75	86	86	85	85	84	71	70	73	74	76
SVM	80	79	77	78	74	91	91	90	89	88	82	81	81	81	78
ANN	78	77	72	71	71	93	92	91	87	83	79	78	77	77	76

International Journal of Applied Mathematics
Volume 38 No. 9s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (online version)

Table 2: Performance evaluation of the best baseline MGC systems with feature selection.

Classifiers		Western					Hindustani					Carnatic				
		Accuracy (%)					Accuracy (%)					Accuracy (%)				
		30s	15s	10s	6s	3s	30s	15s	10s	6s	3s	30s	15s	10s	6s	3s
SVM	XGB	82	81	78	78	76	94	93	91	90	89	82	82	81	81	79
	PCA	81	80	78	78	75	94	93	92	89	88	82	81	81	80	79
ANN	XGB	80	76	78	76	74	95	92	93	89	89	79	78	78	78	78
	PCA	79	79	78	78	75	94	94	93	90	89	81	80	79	79	79

5.2 Deep learning based MGC systems

The CNN model, trained directly on mel-spectrograms, surpassed the traditional models, confirming the advantage of automatic feature learning. However, the proposed CNN-LSTM model achieved the highest overall accuracy, proving its ability to capture both spectral and temporal patterns.

The performances of the CNN models and the proposed CNN-LSTM models are given in Table 3 which demonstrate the superiority of the CNN-LSTM model over the CNN-only approach. To further analyze how the models behaved, we examined confusion matrices along with other evaluation metrics. The confusion matrices for the proposed CNN-LSTM models on Western, Hindustani, and Carnatic datasets for 3 second signals are shown in Figure 3, Figure 4, and Figure 5, respectively.

Table 3: Performance evaluation of CNN and CNN-LSTM MGC systems.

Classifiers	Western					Hindustani					Carnatic				
	Accuracy (%)					Accuracy (%)					Accuracy (%)				
	30s	15s	10s	6s	3s	30s	15s	10s	6s	3s	30s	15s	10s	6s	3s
CNN	83.05	83.92	85.25	86.13	87.05	95.43	96.75	97.85	98.17	98.78	85.32	87.82	90.50	93.05	97.35
CNN-LSTM	84.55	85.22	87.32	88.45	89.63	95.89	97.39	98.05	98.69	99.09	87.50	92.19	95.54	96.70	97.95

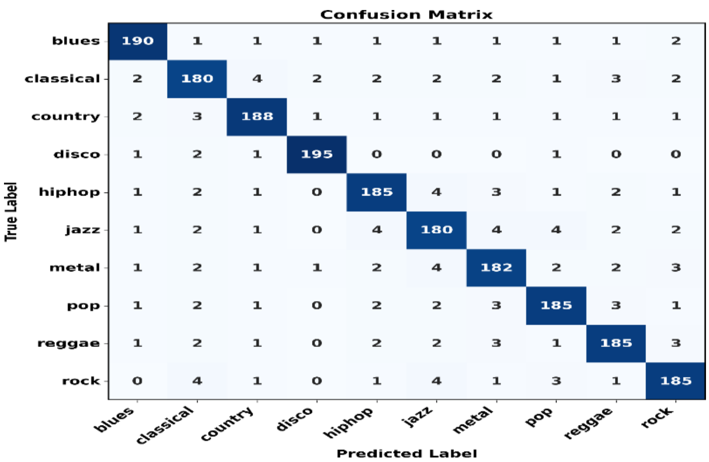

Figure 3: Confusion matrix of CNN-LSTM-based MGC system (3 seconds) Western genre.

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (online version)

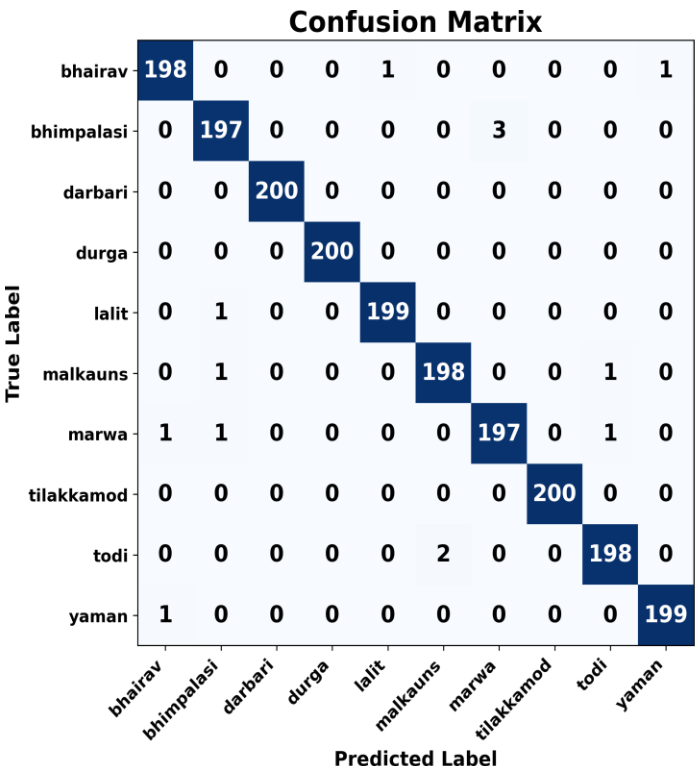

Figure 4: Confusion matrix of CNN-LSTM-based MGC system (3 seconds) Hindustani genre.

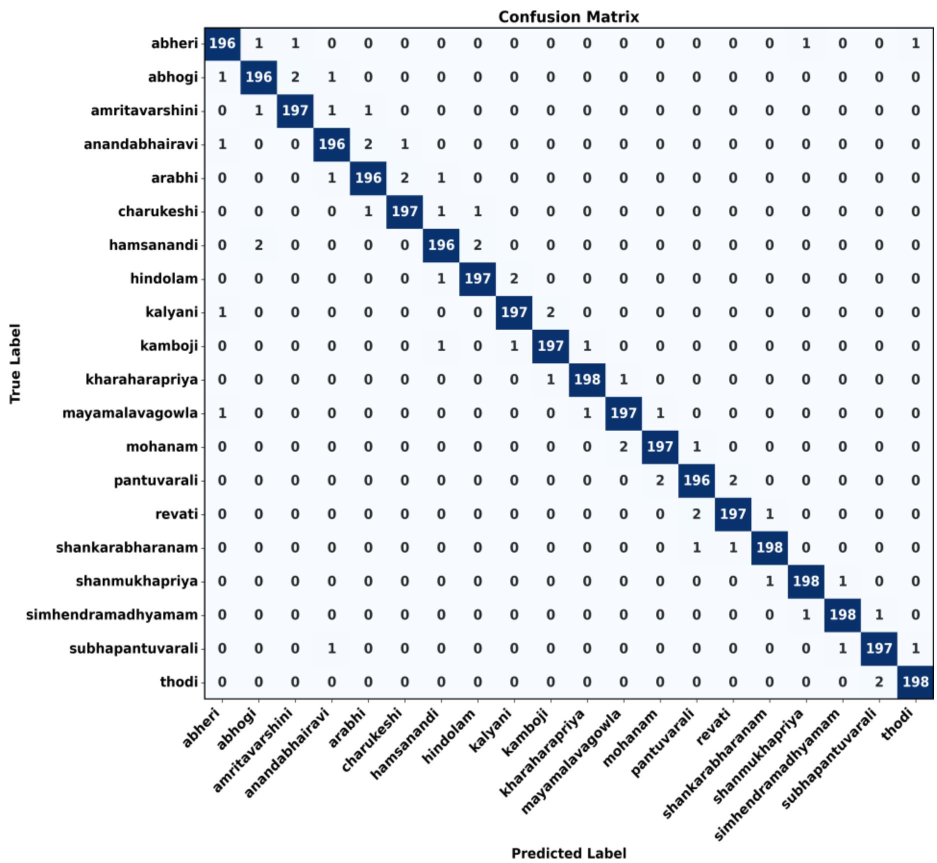

Figure 5: Confusion matrix of CNN-LSTM-based MGC system (3 seconds) Carnatic genre.

The findings from this work underline the importance of integrating both spectral and temporal modeling for music classification across cultures. Traditional machine learning techniques such as

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (online version)

SVM and ANN proved effective when combined with well-crafted audio features, but their performance reached a limit compared to deep learning methods. CNN-based architectures, by contrast, were able to learn feature hierarchies directly from spectrograms, eliminating the need for extensive manual feature engineering. The hybrid CNN-LSTM model took this a step further by capturing the time-evolving nature of music, which is especially valuable for Indian classical styles characterized by improvisation and gradual melodic development. Another key insight is the influence of segment duration. Longer audio segments generally provide richer contextual information, improving accuracy; however, the CNN-LSTM model maintained notable robustness even for short clips, making it suitable for practical applications like mobile based music recognition or streaming recommendation systems. Our results also emphasize the need for more culturally inclusive datasets. While GTZAN remains a standard benchmark, the inclusion of curated Hindustani and Carnatic datasets provides a broader, more realistic test bed for cross-cultural models. This step moves MIR research toward a more global understanding of musical diversity.

6 Conclusion

This study presented a comprehensive comparison of machine learning and deep learning techniques for music genre classification across Western, Hindustani, and Carnatic traditions. We introduced two curated Indian datasets and evaluated multiple models, from classical classifiers like SVM and ANN to modern deep learning architectures. The proposed CNN-LSTM hybrid model achieved the best overall results, effectively learning both the tonal texture and temporal evolution of music. We achieved classification accuracies of 89.63%, 99.09%, and 97.95% for Western, Hindustani, and Carnatic genres, respectively. Our experiments revealed that temporal modeling significantly enhances classification, particularly for complex musical structures like ragas. Even when limited to short clips, the CNN-LSTM model maintained high accuracy, highlighting its potential for real-world applications where data duration is constrained. Future work will explore attention-based and transformer models for richer temporal representation and the use of self-supervised learning to use unlabeled musical data. We also plan to release the curated Indian datasets publicly, enabling further research into culturally diverse music understanding.

References

[1] Hareesh Bahuleyan, *Music genre classification using machine learning techniques*, arXiv preprint arXiv:1804.01149 (2018).

[2] Angelo Cesar Mendes da Silva, Maur´ıcio Archanjo Nunes Coelho, and Raul Fonseca Neto, *A music classification model based on metric learning applied to mp3 audio files*, Expert Systems with Applications **144** (2020), 113071.

[3] MG Deepika and P Sarika, *A comparative analysis of mfis in india using anova and logistic regression model*, International Conference on Artificial Intelligence and Data Engineering, Springer, 2019, pp. 503–515.

[4] Aditya Dhall, YV Srinivasa Murthy, and Shashidhar G Koolagudi, *Music genre classification with convolutional neural networks and comparison with f, q, and mel spectrogram-based images*, Advances in Speech and Music Technology, Springer, 2021, pp. 235–248.

[5] Sakinat O Folorunso, Sulaimon A Afolabi, and Adeoye B Owodeyi, *Dissecting the genre of nigerian music with machine learning models*, Journal of King Saud University-Computer and Information Sciences **34** (2022), no. 8, 6266–6279.

[6] Udaya Baskar G, Vishwa Kiran N, Uday Kumar P, and Syama S, *Support vector machine based short term solar power prediction*, 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), 2022, pp. 264–269.

[7] Wang Hongdan, Siti SalmiJamali, Chen Zhengping, Shan Qiaojuan, and Ren Le, *An intelligent music genre analysis using feature extraction and classification using deep learning techniques*, Computers and Electrical Engineering **100** (2022), 107978.

Received: August 07, 2025

[8] Tao Li, Ogihara, and Mitsunori, *Toward intelligent music information retrieval*, IEEE Transactions on Multimedia **8** (2006), no. 3, 564–574.
[9] Caifeng Liu, Lin Feng, Guochao Liu, Huibing Wang, and Shenglan Liu, *Bottom-up broadcast neural network for music genre classification*, Multimedia Tools and Applications **80** (2021), 7313–7331.
[10] Tamilarasi Kathirvel Murugan and Anush Kameswaran, *Employing convolutional neural networks and explainable artificial intelligence for the detection of seizures from electroencephalogram signal*, Results in Engineering **24** (2024), 103378.
[11] Yannis Panagakis, Constantine L Kotropoulos, and Gonzalo R Arce, *Music genre classification via joint sparse low-rank representation of audio features*, IEEE/ACM Transactions on Audio, Speech, and Language Processing **22** (2014), no. 12, 1905–1917.
[12] M Pushparajan, KT Sreekumar, KI Ramachandran, and C Santhosh Kumar, *Performance enhancement of raga classification systems using recursive feature elimination*, Third International Conference on Sustainable Computing (Singapore), Springer Nature Singapore, 2022, pp. 533–541.
[13] M. Pushparajan, K.T Sreekumar, K.I Ramachandran, and C Santhosh Kumar, *Data augmentation for improving the performance of raga (music genre) classification systems*, 5th International Conference on Electronics and Sustainable Communication Systems (ICESC), IEEE, 2024, pp. 1407–1412.
[14] Jaime Ram´ırez and M Julia Flores, *Machine learning for music genre: multifaceted review and experimentation with audioset*, Journal of Intelligent Information Systems **55** (2020), no. 3, 469–499.
[15] Xi Shao, Changsheng Xu, and Mohan S Kankanhalli, *Unsupervised classification of music genre using hidden markov model*, 2004 IEEE International Conference on Multimedia and Expo (ICME), vol. 3, 2004, pp. 2023–2026.
[16] George Tzanetakis and Perry Cook, *Musical genre classification of audio signals*, IEEE Transactions on speech and audio processing **10** (2002), no. 5, 293–302.
[17] Weibin Zhang, Wenkang Lei, Xiangmin Xu, and Xiaofeng Xing, *Improved music genre classification with convolutional neural networks.*, Interspeech, 2016, pp. 3304–3308.