Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Acoustic Feature Synergy and Self-Supervised Learning for Robust Tabla Stroke Classification

Jaipreet Kaur
Guru Nanak Dev University
Rajdeep Singh Sohal
rajdeep.ece@gndu.ac.in

Guru Nanak Dev University
Manbir Kaur
Thapar Institute of Engineering and Technology
Satinder Kaur
Guru Nanak Dev University

Research Article

Keywords: tabla stroke classification, multidomain feature extraction, handcrafted features, self-supervised learning, ANOVA feature selection

Posted Date: March 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9043295/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Acoustic Feature Synergy and Self-Supervised Learning for Robust Tabla Stroke Classification

Jaipreet Kaur[1], Rajdeep Singh Sohal[1,*], Manbir Kaur[2], Satinder Kaur[3]
[1]Department of Electronics Technology, Guru Nanak Dev University, Amritsar, Punjab, India.
[2]Department of Electrical and Instrumentation Engineering, Thapar Institute of Engineering and Technology, Patiala, Punjab, India.
[3]Department of Computer Engineering & Technology, Guru Nanak Dev University, Amritsar, Punjab, India.

Corresponding Author email: rajdeep.ece@gndu.ac.in

Abstract: Accurate and robust automatic classification of tabla strokes is essential for music information retrieval and performance analysis, yet remains challenging due to complex timbral structures and subtle acoustic variability across stroke categories. To address this challenge, we propose a robust tabla stroke classification framework that integrates multidomain handcrafted features, spanning spectral, temporal, cepstral, and perceptual descriptors together with self-supervised learning (SSL) representations derived from a newly developed, manually annotated tabla dataset. This dataset is accompanied by an augmented counterpart that simulates realistic acoustic variability, enabling systematic evaluation under domain shift conditions. ANOVA F-test based feature selection is applied to retain the most discriminative attributes, and a range of machine learning classifiers are employed. Experimental results show that multidomain feature synergy significantly improves classification performance, with the Hybrid 8 configuration achieving up to 97.56% accuracy under in-domain evaluation, while SSL representations exhibit superior cross-domain robustness, attaining 94.07% accuracy when trained on original data and tested on augmented data. While handcrafted multidomain features yield near-ceiling accuracy in controlled settings, SSL representations provide stronger resilience to acoustic variability. These findings reveal a trade-off between peak discriminative performance and cross-domain generalization, highlighting the complementary strengths of handcrafted and SSL representations for developing robust and generalizable tabla stroke classification systems.

Keywords: tabla stroke classification; multidomain feature extraction; handcrafted features; self-supervised learning; ANOVA feature selection

1. Introduction

The tabla is a central percussion instrument in Indian classical music, serving as a rhythmic backbone in accompaniment as well as a versatile expressive solo instrument. It maintains the structural essence of the chosen raga (melodic framework) and provides rhythmic patterns in musical compositions alongside other Indian percussion instruments such as pakhwaj, mridangam, dholak, and dholki [1-3]. Owing to its refined tonal quality and versatility, the tabla has emerged as the most prominent solo percussion instrument in North Indian classical music [3]. A tabla set consists of two drums: bayan (a bass drum played with the left hand) and dayan (a treble drum, played with the right hand). Distinctive strokes, known as *bols*, are produced through specific articulatory techniques and are represented by mnemonic syllables encoding both physical gesture and acoustic outcome [4]. The acoustic richness of tabla strokes arises from complex variations in attack transients, harmonic content, decay patterns, and spectral distribution, making timbre a crucial perceptual attribute for stroke differentiation [5]. Thus, timbre becomes a key perceptual cue for distinguishing between strokes, even when they share similar pitch or amplitude [6,7]. In tabla performance, strokes often combine resonant and percussive components, producing long decays and sustained overtones that may overlap with subsequent strokes. Moreover, technically distinct strokes such as TI and TA can exhibit striking acoustic similarity, posing challenges for both human listeners and computational classification systems [6,8].

Early computational approaches to tabla stroke classification relied primarily on handcrafted features such as mel-frequency cepstral coefficients (MFCCs), spectral flux and temporal envelopes, combined with

traditional machine learning (ML) models such as hidden Markov models (HMMs), decision trees, and support vector machines. Gillet and Richard pioneered automatic tabla stroke transcription by modeling stroke spectra using Gaussian mixture models within an HMM framework for ten-category classification [9]. Further, Chordia extended this work to larger and more diverse datasets by employing neural networks and tree-based classifiers using spectral and temporal envelope features [8]. Both studies reported substantial drops in accuracy when tested on unseen tabla instruments, highlighting the challenge of instrument-independent generalization. Subsequent studies sought to improve stroke discrimination through refined feature representations. For example, Deolkar and Abraham demonstrated that carefully designed feature representations and detailed time-frequency analysis could improve classification of overlapping strokes [10], while the authors in [2] conducted a detailed time-frequency and wavelet-based analysis to characterize dominant frequencies and harmonic structures across tabla strokes. Despite these advances, robustness across performers, instruments, and recording environments remained limited.

Beyond stroke-level classification, researchers have explored higher-level analytical tasks in tabla music. A research group proposed a syllable-based framework for discovering recurring percussion patterns in tabla solo recordings by combining syllabic transcription with approximate pattern matching [3]. Addressing dataset limitations, Gowiparsad and his team released a large-scale dataset comprising over 38 hours of tabla solo recordings and proposed methods for concert segmentation and gharana recognition [11]. Furthermore, the task of tabla stroke classification also shares conceptual parallels with automatic drum transcription (ADT) in Western popular music, focusing on detecting bass drum (BD), snare drum (SD), and hi-hat (HH) events. Tzanetakis and Cook introduced MARSYAS, a framework for audio content analysis that combined feature extraction, classification, and a novel texture-based temporal segmentation method [12]. Additionally, studies by [13] and [4] established robust ADT pipelines that later inspired adaptations for Indian classical percussion. Leveraging these connections, the research group in [4] mapped tabla stroke categories (bass, treble, and damped strokes) to Western drum counterparts and applied CNN-based architectures to achieve improved four-way tabla stroke classification.

More recently, deep learning has enabled significant progress across a range of tabla-related tasks. The researchers in [14] proposed deep-learning-based frameworks for tabla music generation by training advanced LSTM variants and transformer models, extending insights gained from piano music generation to the domain of classical tabla music. Similarly, the authors in [15] introduced a meta-learning framework for tabla stroke and tala recognition under limited annotated data, incorporating synthetic data generation to improve generalization.

Despite these advances in tabla-related audio analysis, many existing approaches rely on relatively small datasets, lack systematic evaluation under domain shift, or do not explicitly assess robustness across instruments and recording environments. As a result, models that perform well under controlled conditions often degrade significantly when applied to real-world recordings. A fundamental challenge underlying these limitations is the domain shift problem, where models trained on clean, controlled audio fail to generalize to acoustically varied conditions. In tabla stroke classification, such variability arises from differences in microphone placement, room acoustics, performer articulation, and instrument tuning, all of which can substantially alter the acoustic characteristics of strokes. Addressing this challenge is essential for deploying reliable transcription and analysis systems in practical, real-world settings. Recent studies have demonstrated that controlled audio augmentation provides an effective means of simulating realistic recording variability while preserving the semantic integrity of class labels [16,17], which remain largely unexplored for tabla stroke classification.

Motivated by the musicological importance of tabla stroke categorization and the limitations of prior work, this study focuses on robust tabla stroke classification by integrating multidomain handcrafted features

spanning spectral, temporal, cepstral, and perceptual descriptors with learned representations derived from machine learning models. To support this investigation, we curate a labeled dataset of tabla solo recordings and construct an augmented counterpart that emulates realistic acoustic variability while preserving stroke identity. This dual-domain design enables systematic evaluation of both in-domain accuracy and cross-domain robustness, offering deeper insights into the strengths and limitations of different feature representations and learning paradigms.

The key contributions of this work are summarized as follows:

- A new manually annotated tabla stroke dataset comprising nine distinct stroke categories (DHA, GEGE, GEGETIT, GHE, KE, KITTAK, TAA, TINA, TIT), released as 30-second audio segments with stroke-level labels and metadata. The dataset is publicly available at: https://github.com/RSSohal/Tabla-Strokes-Classification to promote reproducible research.
- A multidomain feature-synergy framework integrating spectral, temporal, cepstral, and perceptual descriptors to model the multidimensional timbral structure of tabla strokes.
- A robust experimental framework under simulated domain shift, emulating realistic recording variability to evaluate cross-domain generalization across diverse acoustic conditions.
- Adaptation of self-supervised representation learning framework to learn domain-invariant acoustic representations directly from raw waveform inputs, enhancing robustness to recording variability.
- Comprehensive evaluation across a range of ML classifiers, providing detailed insights into the trade-offs between handcrafted feature-based and representation-learning approaches in terms of classification accuracy and generalization capability.

The remainder of this paper is organized as follows. Section 2 presents the proposed methodology, including signal acquisition and dataset description, simulated domain shift construction, preprocessing steps, multidomain handcrafted feature extraction (spectral, cepstral, temporal, and perceptual), feature selection strategy, self-supervised representation learning framework, and downstream classification methodology. Section 3 presents the experimental results and discussion, covering in-domain evaluation, cross-domain robustness analysis, and comparative performance assessment across feature representations. Section 4 concludes the paper and outlines directions for future research.

2. Methodology

This study proposes a comprehensive feature-synergy framework for the classification of tabla bols using machine learning classifiers. The methodology integrates spectral, temporal, cepstral, and perceptual features extracted from high-quality audio recordings, enabling a richer representation of the unique acoustic signatures of tabla strokes. The overall architecture of the proposed robust tabla stroke classification framework is illustrated in Figure 1, which integrates multidomain handcrafted features and self-supervised learning representations under simulated domain shift.

2.1. Signal Acquisition and Dataset Description

Tabla stroke recordings were collected in a controlled acoustic environment using a high-quality microphone to minimize ambient noise and recording artifacts. All audio signals were recorded at a sampling rate of 16 kHz and stored in mono format to ensure consistency across samples. The dataset comprises nine distinct tabla strokes, selected to capture a broad range of hand techniques, acoustic complexity, and musical usage within the tabla performance tradition. Each stroke represents a unique sonic event characterized by its physical execution technique, acoustic signature, and functional role in Indian classical rhythm.

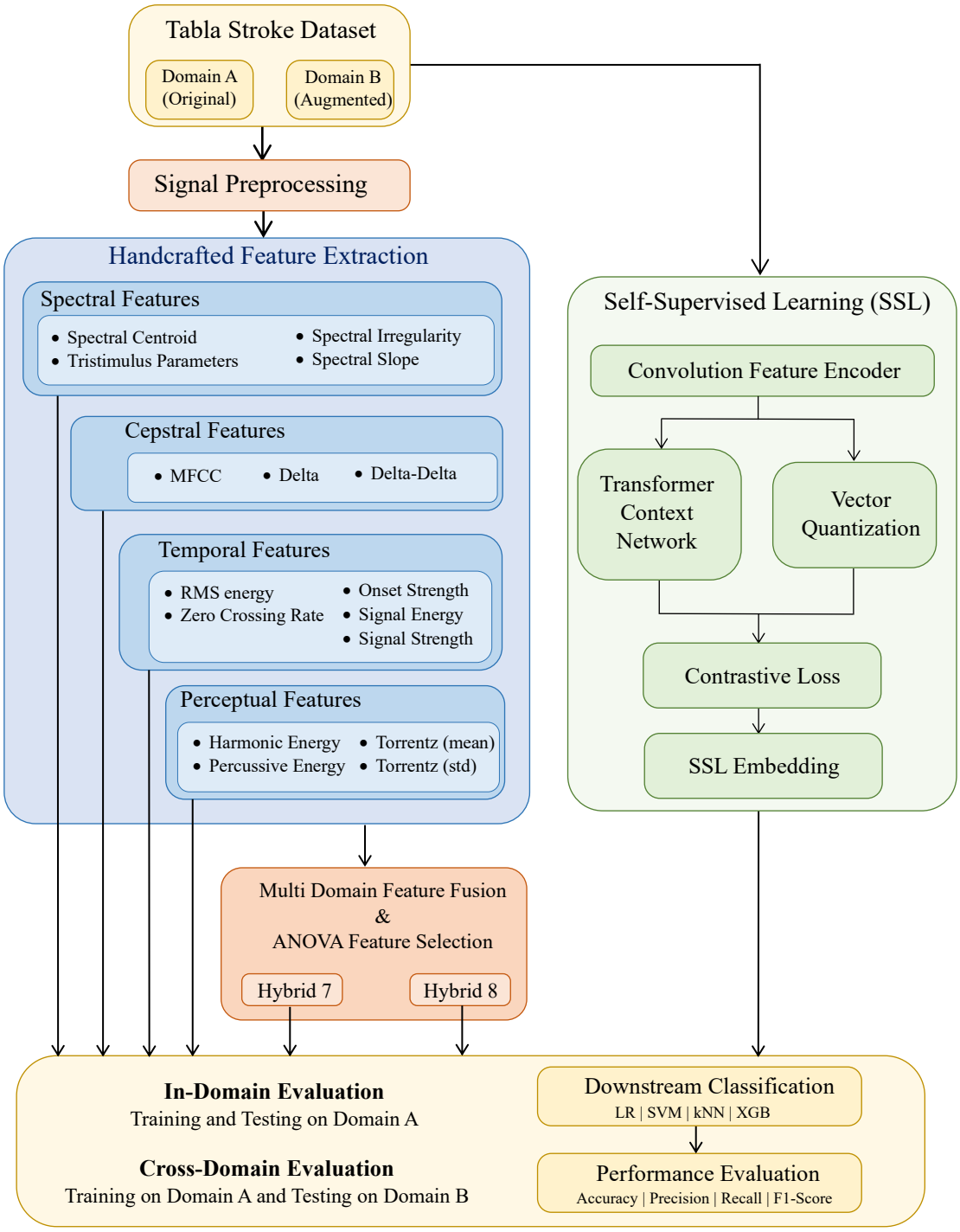


Figure 1: Block diagram of the proposed robust tabla stroke classification framework combining multidomain handcrafted features and self-supervised learning representations under simulated domain shift.

2.1.1. Tabla Strokes Taxonomy

A detailed description of the nine selected tabla strokes is provided in Table 1, which links physical playing gestures to measurable acoustic characteristics and typical musical usage. This taxonomy is compiled from

established musicological literature and prior acoustic analyses of tabla timbre, rather than derived from a single source [2,8,9].

Table 1: Tabla stroke taxonomy with execution technique, acoustic characteristics and musical usage.

Stroke	Execution Technique	Acoustic Characteristics	Musical Usage
DHA	Simultaneous NA (dayan) and GE (bayan) strikes, NA is executed on the dayan using the index or middle finger near the edge, while GE is produced on the bayan using heel pressure and sliding motion.	Rich overtones, high amplitude, dual frequency (bass and treble) components	Found on sam (first beat) of a taal, used in Teentaal, Jhaptal; common in kayda and tukra
GEGE	Two rapid GE strokes played consecutively on bayan using heel press	Low frequency, moderate resonance, pulse-like temporal spacing	Used in *relas*, *parans*, and *kayda* variations to create a rolling bass effect.
GEGETIT	Two GE strokes (bayan) followed by two TI strokes (dayan)	Tonal contrast (bass vs treble), discrete onsets, rhythmically engaging	Appears in improvisations and tension-building phrases
GHE	Heel press and glide heel on bayan after strike	Deep bass, long decay, pitch-shifting due to pressure variation	Common in slow-tempo compositions; provides bass tone
KE	Muted strike on bayan using fingertips	Short duration, low resonance	Used as rhythmic filler in phrases like DHA, KE and TAA
KITTAK	Composite stroke combining KI (closed bayan stroke) and TAK (sharp dayan strike)	Bright timbre, fast attack and rapid decay	Widely used in tukras, parans, and relas for dynamic flair
TAA	Open strike on rim of dayan using fingers	Ringing treble tone, high frequency energy, moderate sustain	Found in light classical genres (Dadra, Keharwa); adds rhythmic texture
TINA	Rapid sequence of TIN (shyahi strike) followed by NA (rim strike)	Two-peaked waveform, resonant onset followed by sharp decay	Used in transitions and fillers to create tonal contrast
TIT	Repeated fingertip taps on shyahi of the dayan	Very short, high-pitched strokes; fluttering effect in rapid repetition	Integral to relas and fast tukras; builds rhythmic intensity

2.1.2. Time-Domain Analysis

Figure 2 illustrates the time-domain waveforms of nine representative tabla strokes, highlighting distinct temporal envelopes, attack–decay behavior, and amplitude dynamics that directly reflect the underlying playing techniques. Bass-dominant strokes such as GHE and GEGE, produced on the bayan, exhibit smoother amplitude envelopes with longer decay times, corresponding to sustained low-frequency resonance and pressure-based excitation. In contrast, treble-oriented strokes such as TIT and KE, executed on the dayan, are characterized by short-duration, high-energy transients with sharp attack and rapid decay, indicative of muted or fingertip-based strikes. Composite strokes, including DHA, GEGETIT, KITTAK, and TINA, display more complex waveform structures, combining rapid onset components with variable decay profiles arising from the interaction of bass and treble components.

2.1.3. Time-Frequency Analysis

Figure 3 presents the power spectral density (PSD) plots (left) and corresponding spectrograms (right) of the same nine strokes. The PSD analysis shows that bass-dominant strokes concentrate most of their energy in the low-frequency region, while treble-oriented strokes exhibit energy concentrated in higher frequency bands with sharper spectral peaks and faster attenuation. Composite strokes demonstrate more complex spectral distributions that combine both low- and high-frequency components. The spectrograms further

reveal distinctive temporal evolution of harmonic content and decay patterns that closely align with the physical excitation mechanisms of each stroke [18,19].

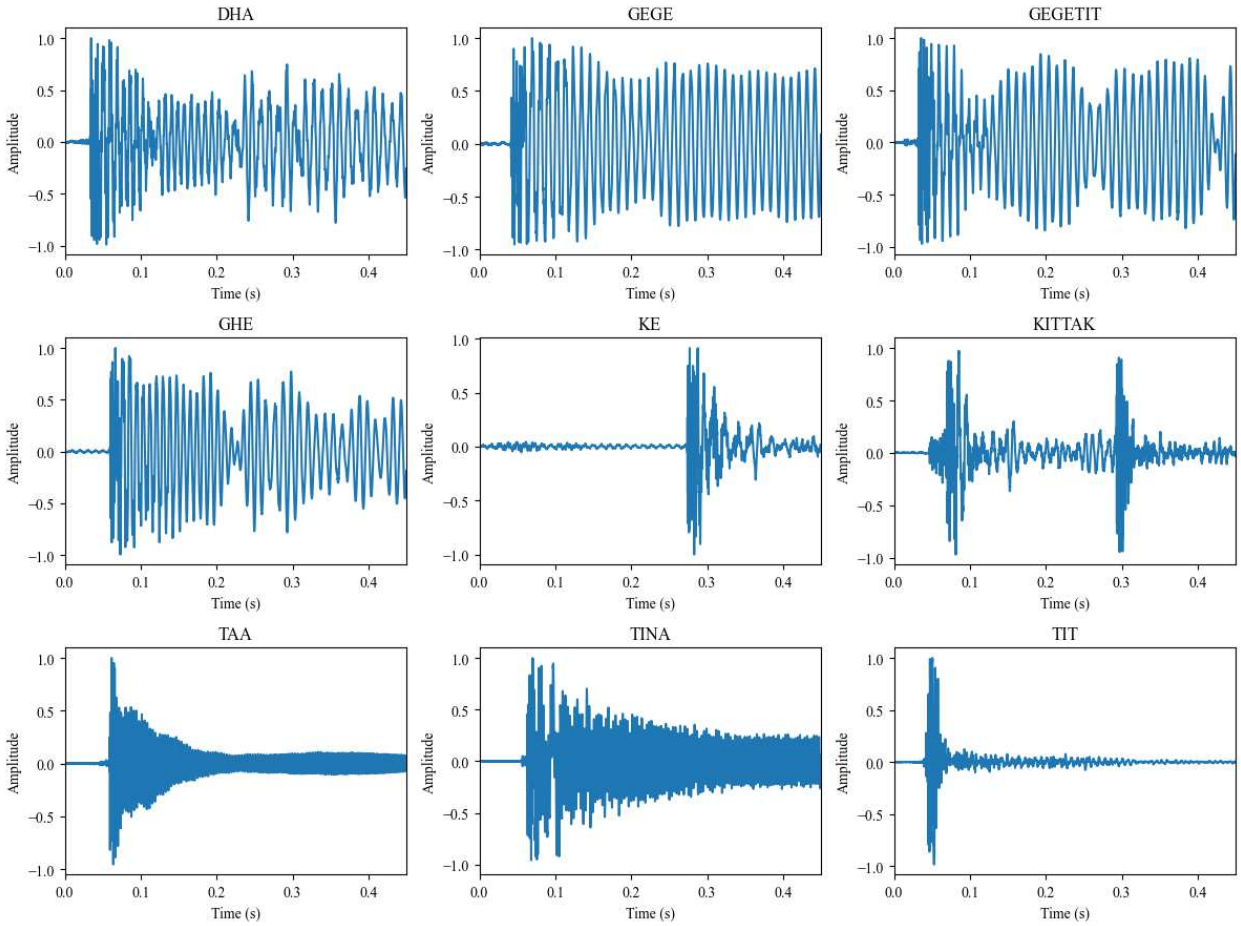
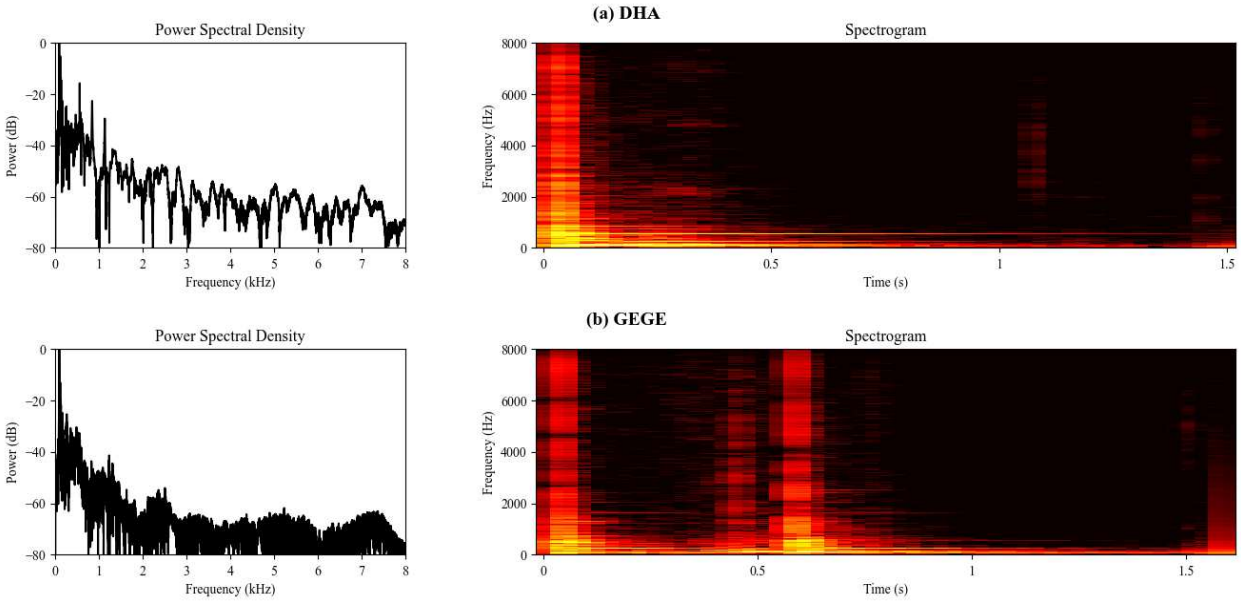
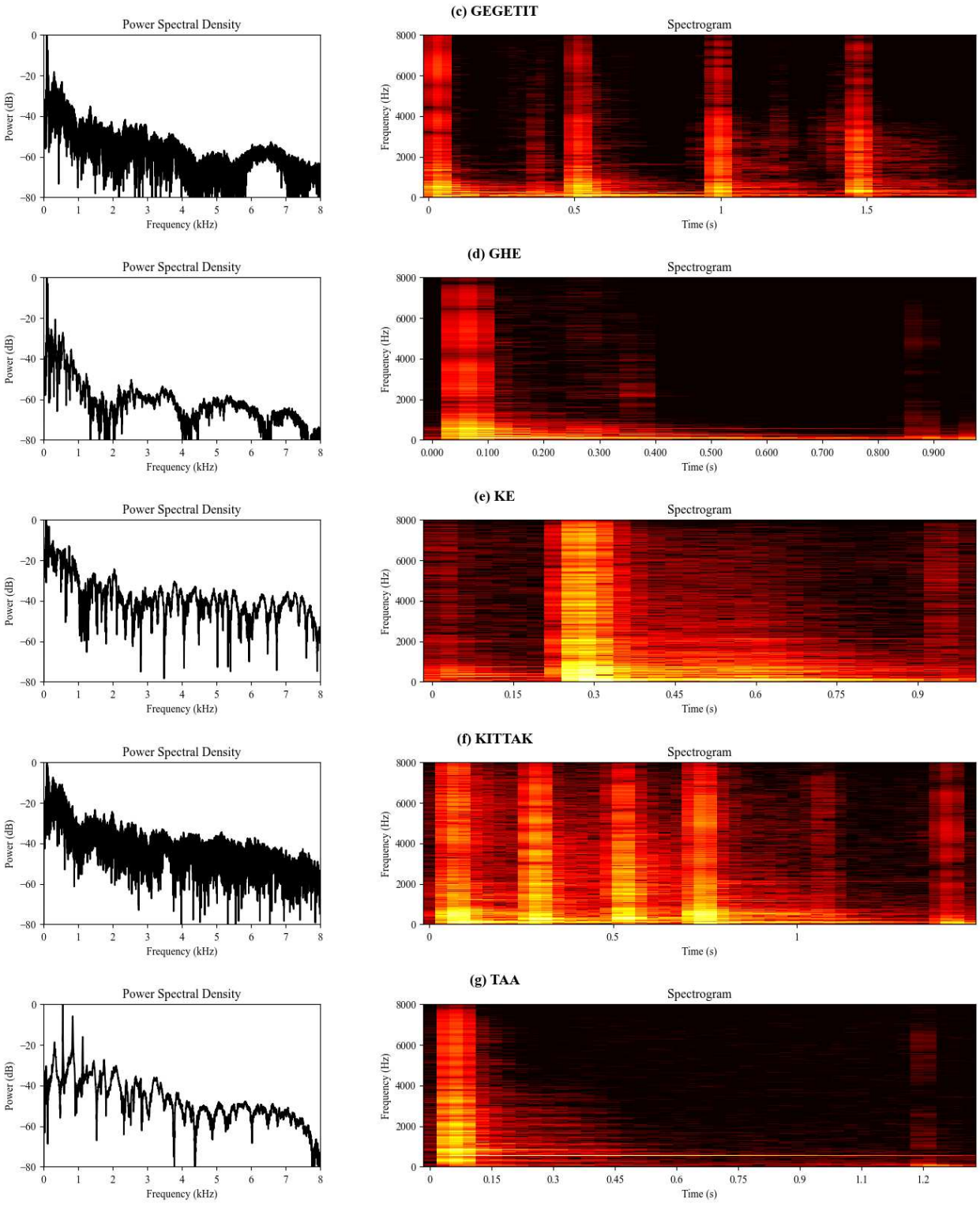


Figure 2: Time-domain waveforms of nine tabla strokes: (a) DHA, (b) GEGE, (c) GEGETIT, (d) GHE, (e) KE, (f) KITTAK, (g) TAA, (h) TINA, and (i) TIT.




(c) GEGETIT
Power Spectral Density
Power (dB)
Frequency (kHz)
Spectrogram
Frequency (Hz)
Time (s)
(d) GHE
Power Spectral Density
Power (dB)
Frequency (kHz)
Spectrogram
Frequency (Hz)
Time (s)
(e) KE
Power Spectral Density
Power (dB)
Frequency (kHz)
Spectrogram
Frequency (Hz)
Time (s)
(f) KITTAK
Power Spectral Density
Power (dB)
Frequency (kHz)
Spectrogram
Frequency (Hz)
Time (s)
(g) TAA
Power Spectral Density
Power (dB)
Frequency (kHz)
Spectrogram
Frequency (Hz)
Time (s)

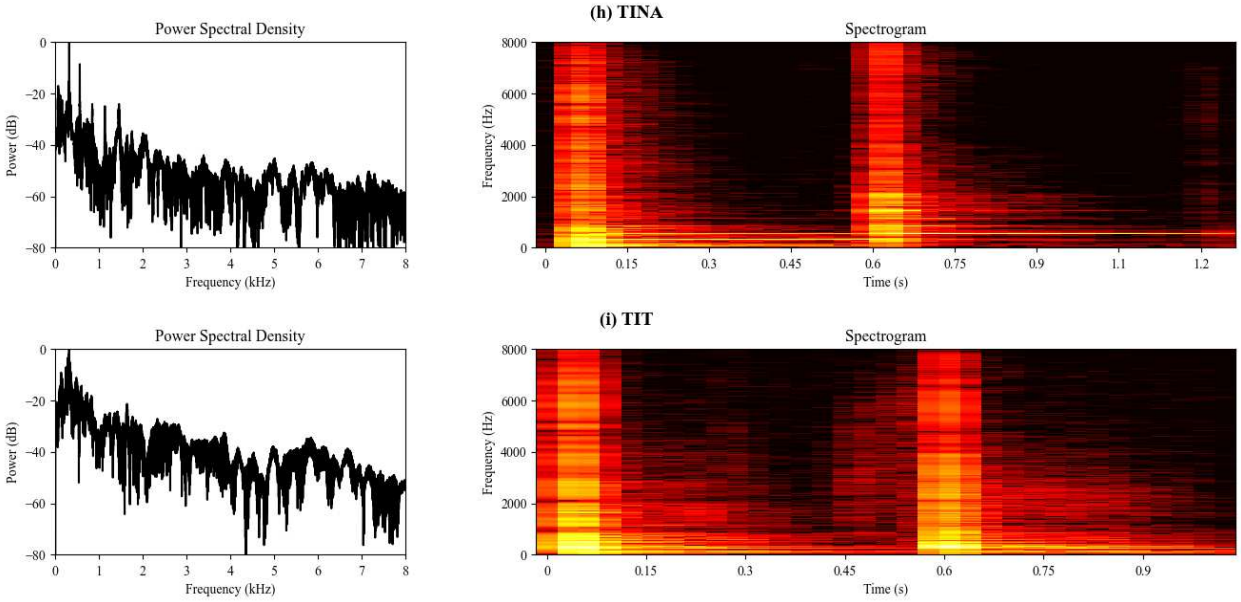


Figure 3: Representative power spectral density and spectrograms of the nine tabla strokes: (a) DHA, (b) GEGE, (c) GEGETIT, (d) GHE, (e) KE, (f) KITTAK, (g) TAA, (h) TINA, and (i) TIT.

Together, the time-domain and time-frequency analyses reveal a rich, multidimensional acoustic structure across tabla strokes, motivating the integration of both handcrafted features and learned representations for robust automatic tabla stroke classification.

2.1.4. Simulated Domain Shift Dataset

Controlled audio augmentation has emerged as an effective strategy for simulating realistic recording variability while preserving the semantic integrity of class labels [16,17]. Motivated by this paradigm, we construct a simulated domain shift dataset to rigorously evaluate the robustness of the proposed tabla stroke classification system under realistic acoustic variations. The original dataset (Domain-A) consists of clean tabla stroke recordings, meticulously organized by stroke category and captured under controlled acoustic conditions. Building on this baseline, an augmented dataset (Domain-B) was generated to emulate realistic acoustic variability commonly encountered in practical recording environments.

Table 2: Configuration of audio augmentations used for simulated domain shift dataset (Domain-B).

Augmentation Type	Parameter Settings	Probability	Simulated Real-World Variation
Gaussian Noise	min_amplitude = 0.0004 max_amplitude = 0.0025	0.6	Microphone self-noise, background environmental noise
Time Stretching	min_rate = 0.98 max_rate = 1.02	0.5	Natural tempo fluctuations during performance
Pitch Shifting	min_semitones = −0.1 max_semitones = +0.1	0.4	Slight tuning differences across instruments and sessions

To preserve the semantic identity of each tabla stroke while introducing controlled variability, a set of perceptually safe augmentations was applied. These include the addition of low-amplitude Gaussian noise to model sensor and background noise, mild time-stretching to reflect natural tempo fluctuations during performance, and minimal pitch shifting to account for slight tuning variations across instruments. The exact parameter ranges and application probabilities are summarized in Table 2. The resulting Domain-B

dataset provides a controlled yet realistic evaluation benchmark for assessing cross-domain generalization under simulated acoustic variability.

2.2. Signal Preprocessing

Figure 1 illustrates the preprocessing pipeline adopted to convert raw tabla audio recordings into time-frequency representations suitable for robust feature extraction and automatic stroke classification [20,21].

Step 1: Pre-Emphasis

Tabla strokes often exhibit a natural spectral tilt, with higher frequency components attenuated relative to lower frequencies. To compensate for this, a first-order pre-emphasis filter is applied to the input signal $x[n]$ to boost high-frequency components. The pre-emphasized sample $y[n]$ is computed as:

$$y[n] = x[n] - \alpha x[n-1]\,, \alpha \in [0.95, 0.99] \tag{1}$$

Step 2: Framing

Although tabla strokes are inherently non-stationary, most spectral and cepstral analysis techniques assume the signal is stationary within the analysis window. Therefore, the pre-emphasized signal is segmented into short, overlapping frames of length M samples (typically 20-40 ms). The framed signal is defined as:

$$y_m[n] = y[n + mH]\ \forall\ n \in \{0,1,\dots,M-1\}\ \&\ m \in \{0,1,\dots,F-1\} \tag{2}$$

where H denotes inter-frame hop size, and F is the total number of frames [21].

Step 3: Windowing

Finite-length framing introduces abrupt discontinuities at frame boundaries, which result in spectral leakage during frequency-domain analysis [22,23]. To mitigate this effect, each frame is multiplied by a Hamming window $w[n]$, which smoothly tapers the frame edges:

$$s_m[n] = y_m[n].\, w[n], \tag{3}$$

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{M-1}\right)\,, n \in [0, M-1] \tag{4}$$

Step 4: Frequency Spectrum Computation

Each windowed frame is transformed into the frequency domain using the Discrete Fourier Transform (DFT). For the mth frame, the complex spectrum is obtained as:

$$S_m[k] = \sum_{n=0}^{M-1} s_m[n] \exp\left(\frac{-j2\pi kn}{M}\right)\,, k \in [0,1,\dots,M-1] \tag{5}$$

The resulting $S_m[k]$ represents the complex spectrum of the mth frame and k indexes the frequency bins, each corresponding to a frequency $f_k = kF_s/M$, where F_s is the sampling frequency [24,25].

The magnitude spectrum, which discards the phase information, is retained as:

$$A_m[k] = |S_m[k]| \tag{6}$$

2.3. Handcrafted Feature Extraction

Tabla strokes generate acoustically rich signals that combine resonant, percussive, harmonic, and transient components depending on the playing technique and drum interaction. Accurately capturing this timbral diversity requires a multi-domain feature extraction framework that integrates complementary information from spectral, cepstral, temporal, and perceptual domains. Following preprocessing, each audio signal is analyzed frame-wise, and multiple feature sets are computed in parallel. These features are subsequently aggregated to form a compact yet expressive representation for classification.

2.3.1. Spectral Features

Spectral features characterize how acoustic energy is distributed across frequencies and provide insight into harmonic structure, brightness, and tone color. These descriptors are particularly effective in differentiating treble-dominant strokes such as TAA, TINA, and TIT, which exhibit stronger high-frequency components, from bass-rich resonant strokes such as GHE, GEGE, and DHA, where acoustic energy is concentrated predominantly in the low-frequency region due to bayan excitation. The most commonly employed spectral features in this study include spectral centroid, tristimulus parameters, spectral slope, and spectral irregularity [26,27].

2.3.1.1. Spectral Centroid

The spectral centroid, SC_m at time frame m represents the "center of mass" of the spectrum and correlates closely with the perceived brightness. Mathematically, it is defined as the weighted mean of harmonic frequencies present in the signal:

$$SC_m = \frac{\sum_{k=1}^{K-1} f_k \,.\, A_m[k]}{\sum_{k=1}^{K-1} A_m[k]} \tag{7}$$

where f_k is the frequency at kth harmonic, $A_m[k]$ is the amplitude of the spectrum at kth harmonic in frame m, K is the total number of harmonics considered (bins).

Higher centroid values indicate dominance of high-frequency harmonics, typical of bright dayan strokes, whereas lower values reflect low-frequency emphasis characteristic of bayan strokes.

2.3.1.2. Tristimulus parameters

Tristimulus descriptors quantify how energy is distributed among the fundamental and harmonic components of a sound:

Tristimulus 1 T_1 quantifies the proportion of energy concentrated in the fundamental frequency:

$$T_1 = \frac{A_1}{\sum_{k=1}^{K} A_k} \tag{8}$$

where A_k is amplitude of the kth harmonic

Among the nine tabla strokes, TIT and TAA typically exhibit higher T_1 values due to their sharp, focused excitation on the dayan.

Tristimulus 2 T_2 represents the proportion of energy in the second and third harmonics:

$$T_2 = \frac{A_2 + A_3}{\sum_{k=1}^{K} A_k} \tag{9}$$

This captures low-order resonance behavior common in strokes such as DHA and GEGETIT.

Tristimulus 3 T_3 aggregates energy from the fourth harmonic onward:

$$T_3 = \frac{\sum_{k=4}^{K} A_k}{\sum_{k=1}^{N} A_k} \tag{10}$$

Higher T_3 values indicate increased spectral complexity and brightness arising from rich overtone content. Among the nine strokes, KITTAK and GEGETIT typically exhibit elevated T_3 values due to their composite nature, involving rapid alternation or combination of bass and treble strokes.

2.3.1.3. Spectral Slope

Spectral slope quantifies the rate of spectral energy decay with frequency and is computed via linear regression on the log-magnitude spectrum:

$$SS_m = \frac{K\sum_{k=1}^{K} f_k.\log\left(A_m[k]\right) - \sum_{k=1}^{K} f_k.\sum_{k=1}^{K}\log A_m[k]}{K\sum_{k=1}^{K} f_k^2 - (\sum_{k=1}^{K} f_k)^2} \tag{11}$$

A flatter spectral slope is observed in treble-dominant strokes such as TAA, TIT, and TINA, reflecting strong high-frequency energy from sharp dayan excitation. In contrast, a steeper negative spectral slope characterizes bass-rich strokes such as GHE, GEGE, and DHA, where low-frequency resonance from the bayan dominates.

2.3.1.4. Spectral Irregularity

Spectral Irregularity quantifies the degree of amplitude fluctuation between adjacent harmonics, providing insight into the roughness or degree of fluctuation of a sound [28]:

$$SI_m = \frac{\sum_{k=1}^{K-1}(A_m[k] - A_m[k+1])^2}{\sum_{k=1}^{K} A_m[k]^2} \tag{12}$$

High spectral irregularity is characteristic of composite strokes such as DHA, GEGETIT, KITTAK, and TINA, where mixed bayan-dayan excitation introduces fluctuating harmonic amplitudes. In contrast, harmonically cleaner strokes like TAA, TIT, and KE exhibit low irregularity due to stable, controlled excitation on the dayan. Bass-dominant strokes such as GHE and GEGE show moderate irregularity, reflecting sustained low-frequency resonance with minor pressure-induced variations on the bayan.

2.3.2. Cepstral Features

Cepstral features provide a compact representation of the short-time spectral envelope, closely aligned with human auditory perception [29-31]. In this study, mel-frequency cepstral coefficients (MFCCs) are employed as the primary cepstral descriptors, along with their first-order (delta) and second-order (delta-delta) derivatives, which capture the temporal dynamics and acceleration of spectral changes across successive frames.

2.3.2.1. MFCC

MFCCs provide a compact representation of the short-time spectral envelope of an audio frame. Originally introduced in speech processing by Davis and Mermelstein [29], MFCCs have since been widely adopted in music information retrieval due to their strong alignment with human auditory perception. In tabla stroke analysis, MFCCs effectively capture timbral structure of each stroke, allowing discrimination between resonant, bass-heavy sounds from sharp, high-frequency strokes.

The power spectrum of the mth frame is obtained from the magnitude spectrum as:

$$P_m[k] = \frac{|S_m[k]|^2}{M} \tag{13}$$

Mel filter bank energies are then computed as:

$$E_m[p] = \sum_k P_m[k] \times H_p[k] \tag{14}$$

where $H_p[k]$ is the pth triangular Mel filter response, defined as

$$H_p[k] = \begin{cases} 0 \quad , & f_k < f_{p-1} \; or \; f_k > f_{p+1} \\ \dfrac{f_k - f_{p-1}}{f_p - f_{p-1}}, & f_{p-1} \le f_k \le f_p \\ \dfrac{f_{p+1} - f_k}{f_{p+1} - f_p}, & f_p \le f_k \le f_{p+1} \end{cases} \tag{15}$$

where, f_k denotes the frequency corresponding to the kth FFT bin, f_p, f_{p-1} and f_{p+1} denote the center, lower, and upper frequencies of the pth Mel filter, respectively.

As human perception of loudness follows an approximately logarithmic scale, the Mel filter bank energies are logarithmically compressed:

$$\hat{E}_m[p] = \log(E_m[p]) \tag{16}$$

To decorrelate the log Mel energies and obtain the cepstral representation, a Discrete Cosine Transform (DCT) is applied:

$$C_m[c] = \sum_{p=1}^{P} \hat{E}_m[p] \cos\left[\frac{\pi c}{P}\left(p - \frac{1}{2}\right)\right] \; \forall \; c \in [1,2, \dots, C] \tag{17}$$

where $C_m[c]$ is the cth cepstral coefficient of frame m, P is the number of Mel filters, and C is the number of retained cepstral coefficients.

2.3.2.2. Delta Coefficients

Delta coefficients capture the first-order temporal dynamics of MFCCs by modeling their rate of change across consecutive frames. They are computed as:

$$\Delta C_m[c] = \frac{\sum_{q=1}^{N} q\left(C_{m+q}[c] - C_{m-q}[c]\right)}{2 \times \sum_{q=1}^{N} q^2} \tag{18}$$

2.3.2.3. Delta-Delta Coefficients

Delta-delta coefficients represent the second-order temporal derivatives of the MFCCs, capturing the acceleration of spectral changes over time:

$$\Delta^2 C_m[c] = \frac{\sum_{q=1}^{N} q\left(\Delta C_{m+q}[c] - \Delta C_{m-q}[c]\right)}{2 \times \sum_{q=1}^{N} q^2} \tag{19}$$

2.3.3. Temporal Features

Temporal features capture time-varying characteristics of tabla strokes, reflecting the rhythm, amplitude fluctuations, and onset patterns inherent in percussive performances. These features complement spectral and cepstral descriptors by providing information about the evolution of signal energy and structure over time [32, 33]. The following temporal descriptors are extracted from each frame of the pre-processed signal.

2.3.3.1. Root Mean Square Energy

Root mean square (RMS) energy measures the average signal amplitude within a frame and serves as a robust indicator of stroke intensity. For the mth frame, RMS energy is computed as:

$$E_{RMS}(m) = \sqrt{\frac{1}{M}\sum_{n=0}^{M-1} s_m[n]^2} \tag{20}$$

Higher RMS values correspond to energetically strong strokes such as DHA and GHE, while lower values are typically observed in muted strokes such as KE and TIT.

2.3.3.2. Zero Crossing Rate

The zero-crossing rate quantifies the rate at which the signal changes sign within a frame, reflecting its noisiness or percussiveness. It is defined as:

$$ZCR(m) = \frac{1}{2M}\sum_{n=1}^{M-1} |g(s_m[n]) - g(s_m[n-1])| \tag{21}$$

where, sign function $g(x)$ is given by

$$g(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases} \tag{22}$$

High ZCR values are typically associated with sharp, transient strokes such as TIT and KITTAK, whereas bass-dominant and resonant strokes such as GHE and GEGE exhibit lower ZCR values due to smoother waveform transitions.

2.3.3.3. Onset Strength

Onset strength captures sudden changes in energy or spectral content, corresponding to the attack of a tabla stroke. It is computed as:

$$O(m) = \max(0, E_m - E_{m-1}) \tag{23}$$

where E_m and E_{m-1} represent the frame-wise energies of the current and previous frames, respectively. Strong onset values indicate clearly articulated attacks, which are prominent in strokes such as TAA, TINA, and KITTAK.

2.3.3.4. Signal Energy

Signal energy provides a measure of the total power contained in a frame or the entire signal:

$$E_m = \sum_{n=0}^{M-1} s_m[n]^2 \tag{24}$$

This feature helps distinguish sustained strokes with extended decay, such as GHE and DHA, from short-duration strokes with rapidly diminishing energy, such as KE and TIT.

2.3.3.5. Signal Entropy

Signal entropy quantifies the distribution of energy within a frame, indicating the uniformity or complexity of the stroke. Each frame is divided into S subframes, and the energy of the ith subframe is denoted by $E_{m,i} \forall\, i \in [1, \ldots, S]$. The normalized subframe energy is computed as

$$p_{m,i} = \frac{E_{m,i}}{\sum_{j=1}^{S} E_{m,j}}. \tag{25}$$

The Shannon entropy of the frame is then computed as:

$$H(m) = -\sum_{i=1}^{S} p_{m,i} \log_2\left(p_{m,i}\right) \tag{26}$$

Higher entropy values indicate greater energy dispersion and temporal complexity, commonly observed in composite strokes such as DHA, GEGETIT, and KITTAK, whereas lower entropy values correspond to simpler, more uniform strokes such as TAA and KE.

2.3.4. Perceptual Features

Perceptual features capture characteristics of the audio signal that closely align with human auditory perception, emphasizing aspects such as harmonicity, percussiveness, and tonal relationships [34,35]. These features complement spectral, cepstral, and temporal features by bridging the gap between signal-level analysis and human listening experience.

2.3.4.1. Harmonic energy

Harmonic energy quantifies the total energy contained in the harmonically related frequency components of a tabla stroke. For mth frame, it is computed as:

$$E_h(m) = \sum_{k \in \mathcal{H}} A_m(k)^2 \tag{27}$$

where $\mathcal{H}$ denotes the set of frequency bins corresponding to harmonic components. High harmonic energy values are typically observed in resonant strokes such as GHE, DHA, and TINA, where sustained tonal components dominate the perceptual impression.

2.3.4.2. Percussive energy

Percussive energy measures the energy contained in the broadband and transient components of the signal. It is defined as:

$$E_p(m) = \sum_{k \in \mathcal{P}} A_m(k)^2 \tag{28}$$

where $\mathcal{P}$ denotes the set of frequency bins associated with percussive or non-harmonic components. Strokes such as TIT, KE, and KITTAK, which are executed using quick fingertip or muted strikes, typically exhibit higher percussive energy due to their impulsive nature.

2.3.4.3. Tonnetz (mean)

Tonnetz features describe the tonal and harmonic relationships within the signal by capturing chroma representations into a low-dimensional tonal space. First, chroma extraction maps the harmonic spectrum into 12 pitch-class bins:

$$c_i(m) = \sum_{k \in \mathcal{K}_i} |A_m[k]|^2 \ , i \in [0,1, \dots ,11] \tag{29}$$

The resulting chroma vector for frame m is given by

$$C(m) = [c_0(m), c_1(m), \dots , c_{11}(m)] \tag{30}$$

The Tonnetz transformation further maps the chroma vector into six-dimensional tonal coordinates:

$$T_1(m) = \sum_{i=0}^{11} c_i(m) \cos\left(\frac{7\pi i}{6}\right), \tag{31}$$

$$T_2(m) = \sum_{i=0}^{11} c_i(m) \sin\left(\frac{7\pi i}{6}\right), \tag{32}$$

$$T_3(m) = \sum_{i=0}^{11} c_i(m) \cos\left(\frac{3\pi i}{2}\right), \tag{33}$$

$$T_4(m) = \sum_{i=0}^{11} c_i(m) \sin\left(\frac{3\pi i}{2}\right), \tag{34}$$

$$T_5(m) = \sum_{i=0}^{11} c_i(m) \cos\left(\frac{2\pi i}{3}\right), \tag{35}$$

$$T_6(m) = \sum_{i=0}^{11} c_i(m) \sin\left(\frac{2\pi i}{3}\right). \tag{36}$$

These components are concatenated to form the Tonnetz feature vector:

$$T_d(m) = [T_1(m), T_2(m), T_3(m), T_4(m), T_5(m), T_6(m)] \tag{37}$$

To obtain a stroke-level representation, the frame-wise Tonnetz vectors are aggregated by computing their mean across all F frames:

$$\mu = \frac{1}{F} \sum_{m=1}^{F} T_d(m) \tag{38}$$

The mean Tonnetz vector captures the average tonal organization of a stroke and is particularly informative for strokes with stable harmonic content, such as DHA and GHE.

2.3.4.4. Tonnetz (std)

In addition to the mean, the variability of tonal relationships over time is measured using the standard deviation of the Tonnetz vectors:

$$\sigma_{T_d} = \sqrt{\frac{1}{F}\sum_{m=0}^{F-1}(T_d(m) - \mu)^2} \tag{39}$$

Higher Tonnetz variability is observed in composite or rapidly articulated strokes such as GEGETIT and KITTAK, reflecting dynamic tonal changes, whereas lower variability corresponds to more stable tonal patterns.

A comprehensive overview of timbre domains and corresponding feature groups used for classification is presented in Table 3. The stroke-wise dominance of multidomain acoustic features is summarized in Table 4, highlighting relative feature intensity levels (low, medium, high) across the nine tabla strokes.

Table 3: Timbre domains and feature groups in tabla stroke analysis.

Domain	Focus	Features	Significance of Features
Spectral	Frequency content, harmonic structure [26,27]	1. Spectral centroid	Indicates brightness of sound by locating the "center of mass" of spectral energy.
		2. Tristimulus parameters (T1, T2, T3)	Describes how energy is distributed among the fundamental and harmonic components.
		3. Spectral irregularity	Captures roughness by measuring unevenness between adjacent harmonics.
		4. Spectral slope	Reflects tonal sharpness by quantifying how quickly energy decreases with frequency.
Cepstral	Compact spectral representation; widely used in ML classification [29-31]	1. MFCC	Provides compact spectral envelope information aligned with human auditory perception.
		2. Delta	Captures dynamic changes in cepstral features across frames, aiding rhythm detection.
		3. Delta-Delta	Model acceleration of spectral changes, improving temporal context modeling.
Temporal	Time-based evolution; distinguishes transient vs sustained strokes [32,33]	1. RMS energy	Represents overall signal power, useful for distinguishing stroke intensity.
		2. Zero crossing rate	Measures noisiness or percussiveness in transient strokes.
		3. Onset strength	Detects attack clarity, crucial for transient-rich tabla bols.
		4. Signal energy	Tracks time-domain energy variations to separate sustained from decaying strokes.
		5. Signal entropy	Quantifies unpredictability in amplitude distribution, reflecting stroke complexity.
Perceptual	Auditory descriptors; align computational features with human perception [7,34,35]	1. Harmonic energy	Highlights resonance by measuring tonal energy in harmonics.
		2. Percussive energy	Captures transient, noise-like components typical of sharp tabla strikes.
		3. Tonnetz (mean)	Represents average tonal relations, reflecting perceived harmonic stability.
		4. Tonnetz (std)	Captures variability in tonal relations, useful for subtle perceptual differences between strokes.

Table 4: Dominant multidomain acoustic feature characteristics across nine tabla strokes.

Feature ↓ / Stroke →	DHA	GEGE	GEGETIT	GHE	KE	KITTAK	TAA	TINA	TIT
Spectral Centroid	L	★L	M	★L	★H	★H	★H	★H	★H
Spectral Slope	L	★L	M	★L	★H	★H	★H	★H	★H
Tristimulus T_1	M	M	L	★H	L	L	★H	M	★H
Tristimulus T_2	★H	★H	M	M	L	M	L	M	L
Tristimulus T_3	M	L	★H	L	M	★H	L	★H	L
Spectral Irregularity	★H	M	★H	M	★L	★H	★L	★H	★L
MFCC (static)	★H	★H	M	★H	M	M	M	M	M
Delta / Delta-Delta	M	L	★H	L	★H	★H	M	M	★H
RMS Energy	★H	M	★H	M	★H	★H	M	M	★H
Zero Crossing Rate	M	★L	M	★L	★H	★H	M	M	★H
Onset Strength	★H	M	★H	L	★H	★H	M	M	★H
Signal Entropy	★H	M	★H	L	M	★H	L	M	M
Harmonic Energy	★H	★H	M	★H	L	M	M	M	L
Percussive Energy	M	★L	★H	★L	★H	★H	M	M	★H
Tonnetz (mean)	M	L	★H	★M	L	★H	L	M	★H
Tonnetz (std)	M	L	★H	L	L	★H	L	M	★H

L, M, and H denote low, medium, and high relative feature dominance, respectively. Starred entries (★) indicate stroke-specific discriminative features.

2.4. Multidomain Feature Fusion and Selection

Feature selection plays a crucial role in enhancing performance, improving model generalization capability and reducing overfitting. To evaluate the synergy across domains, four feature groups, namely spectral, temporal, cepstral, and perceptual were merged using the class label as the joining key. This fusion produces a unified multimodal feature space that simultaneously captures spectral, temporal, harmonic, and perceptual characteristics of each stroke, thereby offering a holistic representation of tabla acoustics.

However, the fused representation increases dimensionality and may introduce redundancy. To retain only the most discriminative descriptors, feature selection was performed using the Analysis of Variance (ANOVA) F-test, a statistical method that measures the differences between group means relative to the variability within each group. The ANOVA test computes the F score F_i for each feature x_i, defined as:

$$F_i = \frac{\sigma^2_{between}}{\sigma^2_{within}} \tag{40}$$

where $\sigma^2_{between}$ denotes between-class variance and σ^2_{within} denotes within-class variance.

A higher F score indicates that the feature exhibits greater inter-class variation relative to intra-class variation. Based on descending F-scores, the top-ranked features were identified and retained for downstream classification. As shown in Figure 5, the ANOVA ranking identifies the fifteen highest-scoring features selected for downstream classification. Signal-level descriptors such as Signal Entropy, Signal Energy, and RMS Energy dominate the upper ranks, underscoring the importance of amplitude distribution and temporal energy dynamics in distinguishing tabla strokes. Spectral Irregularity and Harmonic Energy also appear prominently, emphasizing the role of harmonic structure and spectral roughness in stroke differentiation. Several cepstral features, particularly MFCC 2, MFCC 5, MFCC 12, MFCC 13, and MFCC 14, contribute meaningfully, indicating that cepstral coefficients provide complementary spectral envelope information, though with slightly lower discriminative strength compared to signal-level features.

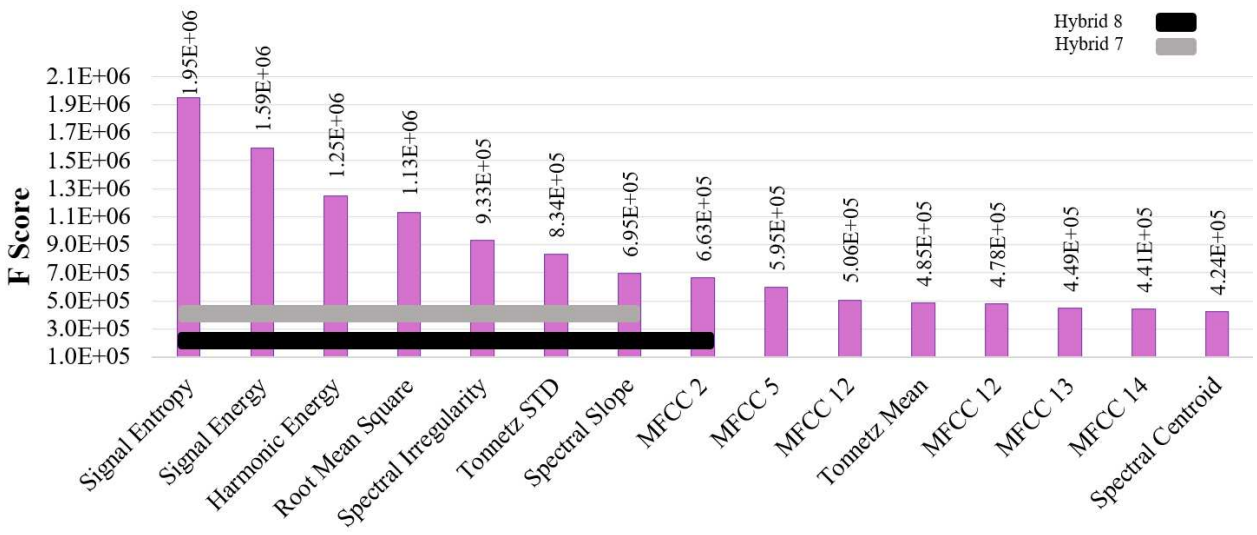


Figure 5: ANOVA F score ranking of the top 15 features used for tabla stroke classification.

To further examine feature synergy, top-ranked descriptors were progressively integrated to form enhanced multimodal subsets. The seven and eight highest-ranked features, denoted as the Hybrid 7 and Hybrid 8 configurations, were selected based on descending ANOVA F scores to ensure maximal discriminative capability with minimal redundancy. By integrating complementary acoustic domains while retaining statistically informative descriptors, the proposed feature-synergy framework improves classification accuracy, reduces computational complexity, and enhances robustness under domain variability.

2.5. Self-Supervised Representation Learning

Tabla acoustics exhibit high variability due to differences in instrument construction, tuning, performer articulation, and recording environments. Supervised models trained on limited labeled data often fail to generalize under such variability. To address this limitation, we adopt a self-supervised representation learning (SSL) framework based on the Wav2Vec 2.0 architecture, depicted in Figure 6, to learn robust waveform level representations without requiring stroke annotations. Unlike handcrafted features (e.g., MFCCs), SSL models learn hierarchical representations by optimizing surrogate objectives that capture temporal and spectral structure. These embeddings are subsequently used as feature representations for downstream stroke classification.

Step 1: Convolution Feature Encoder

Let $x \in \mathbb{R}^T$ denote the raw audio waveform sampled at 16 kHz, where T represents the number of temporal samples. The raw waveform is first processed by a multi-layer temporal convolutional neural network (CNN), which acts as a learnable feature extractor

$$z_t = f_{CNN}(x) \tag{41}$$

where z_t represents latent frame-level embeddings at time index t.

This stage replaces traditional fixed feature extraction pipelines with a fully learnable front-end.

Step 2: Vector Quantization

To enable self-supervised learning, the latent representations z_t are discretized using vector quantization:

$$q_t = Q(z_t) \tag{42}$$

A learned codebook maps continuous latent vectors into discrete acoustic units. These quantized representations q_t serve as stable self-supervised targets during training.

Step 3: Transformer Context Network

The latent representations z_t are randomly masked to produce $\tilde{z_t}$ and passed to the Transformer encoder. This model models long-range temporal dependencies using self-attention mechanisms and produces contextualized embeddings c_t.

$$c_t = f_T(\tilde{z_t}) \tag{43}$$

Because certain time steps are masked, the model is forced to infer missing acoustic content from surrounding context. This masked prediction strategy enables the learning of high-level temporal structure and contextual invariances. Such contextual modeling is particularly important for tabla strokes, where transient attacks and resonant tails overlap temporally.

Step 4: Contrastive Learning Objective

Training is performed using a contrastive loss that encourages the contextual representation c_t to identify the corresponding quantized target q_t among a set of negative samples Q_t:

$$\mathcal{L}_c = -\log \frac{\exp\big(sim(c_t, q_t)\big)}{\sum_{q \in Q_t} \exp\big(sim(c_t, q)\big)} \tag{44}$$

where $sim(\cdot,\cdot)$ denotes cosine similarity.

This objective enforces:

- High similarity between the correct context-target pair (c_t, q_t)
- Low similarity between c_t and incorrect negative samples

As a result, the model learns representations that are both discriminative and invariant to low-level acoustic variability.

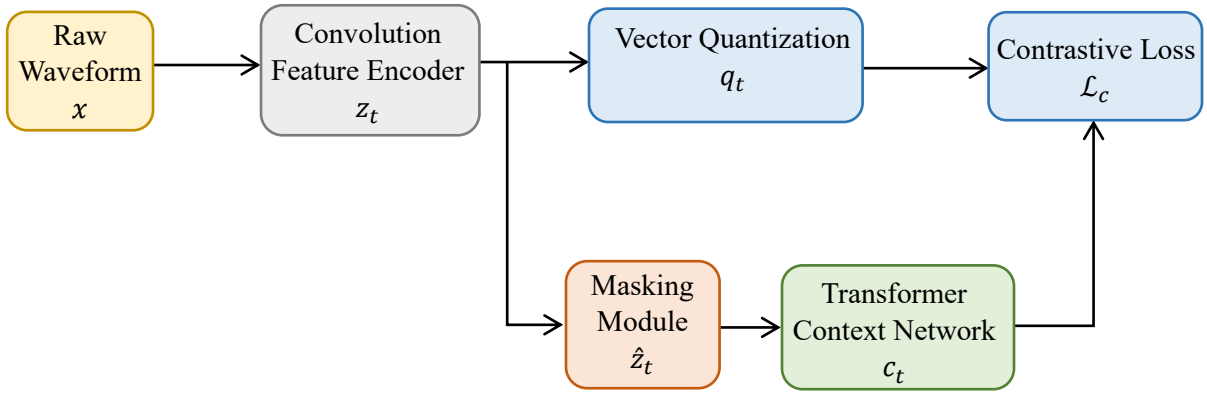


Figure 6: Self-supervised learning pipeline based on the Wav2Vec 2.0 architecture.

2.6. Downstream Classification and Evaluation Metrics

Four widely adopted classifiers were employed, as summarized in Table 5 to assess the separability and robustness of the learned representations. The evaluated classifiers include Logistic Regression (LR), Support Vector Machines (SVM), k-Nearest Neighbors (kNN), and Extreme Gradient Boosting (XGB).

Table 5: Classification Models Used for Downstream Evaluation.

Classifier	Type	Decision Principle	Key Characteristics
Logistic Regression (LR)	Linear model	Softmax-based probabilistic classification	Evaluates linear separability; strong baseline
Support Vector Machine (SVM)	Margin-based model	Maximum margin hyperplane (RBF kernel)	Captures nonlinear decision boundaries
k-Nearest Neighbors (kNN)	Instance-based model	Majority vote of k nearest samples (Euclidean distance)	Exploits local geometric structure
Extreme Gradient Boosting (XGB)	Ensemble tree-based model	Gradient-boosted decision trees	Models nonlinear feature interactions

Hyperparameters were optimized using stratified k-fold cross-validation on the training dataset (Domain-A) to ensure robust model selection and to prevent overfitting. Model tuning was performed exclusively on the training set to avoid information leakage during cross-domain evaluation.

Classification performance was assessed using standard multi-class evaluation metrics, as summarized in Table 6. These metrics provide a comprehensive assessment of predictive performance across all stroke categories.

Table 6: Evaluation Metrics for Multi-Class Classification.

Metric	Formula	Interpretation
Accuracy	$= \frac{TP + TN}{TP + TN + FP + FN}$	Measures overall classification correctness
Precision	$= \frac{TP}{TP + FP}$	Indicates the reliability of positive predictions
Recall	$= \frac{TP}{TP + FN}$	Measures the ability to correctly identify all instances of a class
F1-Score	$= 2.\frac{Precision.Recall}{Precision + Recall}$	Provides a balanced trade-off between precision and recall

where TP, TN, FP and FN denote true positives, true negatives, false positives, and false negatives, respectively.

3. Results and Discussions

This section presents a comprehensive evaluation of tabla stroke classification performance using individual feature domains, hybrid feature combinations, and self-supervised learning representations. The analysis is conducted under two experimental settings:

(i) **In-domain evaluation**: Training and testing on the original dataset (Domain-A)
(ii) **Cross-domain evaluation**: Training on the original dataset (Domain-A) and testing on the augmented dataset (Domain-B) to assess robustness under simulated domain shift.

3.1. In-Domain Evaluation

Table 7 summarizes the classification performance of different feature groups and classifiers when both training and testing are conducted on the original dataset. The corresponding performance metric trends are visualized in Figure 7.

Table 7: Classification performance (%) of different feature groups and classifiers trained and tested on the original (Domain A) tabla stroke dataset.

Feature Group	Model Name	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Spectral	LR	87.8	87.59	88.89	87.49
	SVM	87.8	91.76	88.89	87.65
	kNN	82.93	84.35	83.89	81.68
	XGB	78.05	82.59	79.44	76.84
Cepstral	LR	95.12	95.56	95.56	95.06
	SVM	95.12	95.56	95.56	95.06
	kNN	92.68	93.33	93.33	92.59
	XGB	87.8	90.74	88.33	87.4
Perceptual	LR	85.37	89.84	86.67	84.36
	SVM	80.49	83.76	82.22	80.58
	kNN	80.49	86.67	82.22	79.73
	XGB	85.37	88.52	86.11	86.1
Temporal	LR	90.24	92.59	90.56	90.16
	SVM	92.68	94.44	92.78	92.4
	kNN	90.24	93.39	90.56	90.36
	XGB	82.93	85.93	83.89	82.72
Hybrid 7	LR	95.12	96.3	95.56	95
	SVM	92.68	94.07	93.33	92.53
	kNN	92.68	94.07	93.33	92.53
	XGB	92.68	95.24	93.33	92.96
Hybrid 8	LR	**97.56**	**97.78**	**97.78**	**97.53**
	SVM	**97.56**	**97.78**	**97.78**	**97.53**
	kNN	**97.56**	**97.78**	**97.78**	**97.53**
	XGB	92.68	95.24	93.33	92.96
SSL	LR	95.12	95.56	95.56	95.06
	SVM	95.12	95.56	95.56	95.31
	kNN	92.68	94.07	93.33	92.53
	XGB	85.37	88.89	86.67	85.19

Among the individual feature domains, cepstral features consistently achieve the highest classification performance, with both LR and SVM attaining 95.12% accuracy alongside balanced precision and recall values. This confirms the effectiveness of MFCCs and their temporal derivatives in modeling the short-time spectral envelope of tabla strokes, which closely aligns with human auditory perception of timbre. Temporal features also demonstrate robust performance, achieving up to 92.68% accuracy with SVM, underscoring the importance of onset strength, RMS energy, and temporal energy in distinguishing transient-rich strokes (e.g., TIT and KE) from sustained resonant strokes (e.g., GHE and DHA). However, temporal descriptors alone occasionally confuse strokes with similar rhythmic envelopes, particularly

KITTAK and TINA. Spectral features yield moderate performance, with a maximum accuracy of 87.8%, reflecting their effectiveness in capturing brightness and harmonic distribution. Perceptual features achieve slightly lower accuracy (up to 85.37%), indicating that measures of harmonic and percussive energy provide informative but incomplete representations when used in isolation.

Hybrid feature combinations significantly outperform individual feature domains. The Hybrid 7 feature set achieves up to 95.12% accuracy, while the Hybrid 8 configuration attains the highest overall accuracy of 97.56% across LR, SVM, and kNN classifiers. This near-perfect performance confirms that integrating spectral, temporal, cepstral, and perceptual descriptors yields a richer and more robust representation of tabla timbre. The SSL representations achieve competitive performance of 95.12% accuracy with LR and SVM. While SSL embeddings effectively capture high-level acoustic patterns, they do not surpass the best handcrafted hybrid features under matched training and testing conditions. This suggests that carefully designed multidomain features remain highly effective for structured percussion sounds when training and testing distributions are aligned.

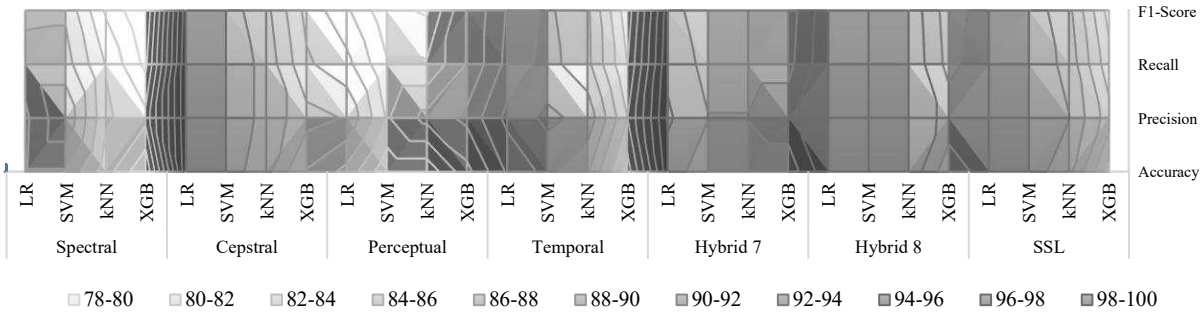


Figure 7: Heatmap visualization of classification metrics across feature domains and classifiers trained and tested on the original tabla dataset.

3.2. Cross-Domain Evaluation

Table 8 reports the performance obtained when models trained on the original dataset (Domain-A) are evaluated on the augmented dataset (Domain-B), providing a stringent assessment of robustness under simulated domain shift. The corresponding performance trends are illustrated in Figure 8. A substantial performance degradation is observed for spectral and temporal features under domain shift, with accuracies dropping to 57-59% for LR, SVM, and kNN. This degradation indicates a high sensitivity of spectral descriptors to additive noise, pitch perturbations, and mild time-stretching, while temporal features suffer from distortions in onset dynamics and amplitude contours. In contrast, cepstral features exhibit markedly stronger robustness, maintaining accuracies of up to 88.89% with kNN. The relative stability of MFCCs under moderate noise and tempo variation explains this behavior, as cepstral representations emphasize perceptually relevant spectral envelopes while suppressing fine-scale fluctuations. Perceptual features demonstrate moderate resilience (up to 85.19% accuracy), suggesting that harmonic-percussive energy measures and tonal descriptors retain useful discriminative power even under acoustic perturbations.

Self-supervised learning representations demonstrate the highest robustness among all feature sets. SSL + LR and SSL + kNN achieve 94.07% and 93.33% accuracy, respectively, outperforming all handcrafted domains in cross-domain evaluation. This result indicates that SSL embeddings capture domain-invariant acoustic structures by modeling global spectral-temporal dependencies rather than relying on narrow class-specific cues. By learning representations from raw waveform structure, SSL suppresses nuisance variability such as background noise, minor pitch shifts, and tempo fluctuations. However, the variability

in SSL performance across classifiers, particularly the reduced accuracy with SVM and XGB, indicates that while SSL features are robust, their ultimate effectiveness depends on the downstream classifier's ability to exploit high-dimensional representations.

Table 8: Classification performance (%) of different feature groups and classifiers trained on the original (Domain A) and tested on the augmented dataset (Domain B) tabla stroke dataset.

Feature Group	Model Name	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Spectral	LR	59.26	77.34	59.26	56.52
	SVM	57.04	67.76	57.04	55.88
	kNN	59.26	64.72	59.26	56.71
	XGB	75.56	77.48	75.56	74.67
Cepstral	LR	85.19	91.93	85.19	85.53
	SVM	84.44	90.85	84.44	84.71
	kNN	88.89	91.73	88.89	89.28
	XGB	65.93	78.35	65.93	67.3
Perceptual	LR	82.22	83.13	82.22	82.16
	SVM	80.74	81.95	80.74	80.65
	kNN	80	81.65	80	80.01
	XGB	85.19	86.78	85.19	85.35
Temporal	LR	68.15	73.82	68.15	68.73
	SVM	67.41	73.07	67.41	68.87
	kNN	63.7	72.81	63.7	65.83
	XGB	71.85	79.19	71.85	72.26
SSL	LR	**94.07**	**95.37**	**94.07**	**93.95**
	SVM	77.04	86.99	77.04	77.81
	kNN	93.33	93.64	93.33	93.26
	XGB	77.78	83.46	77.78	77.42

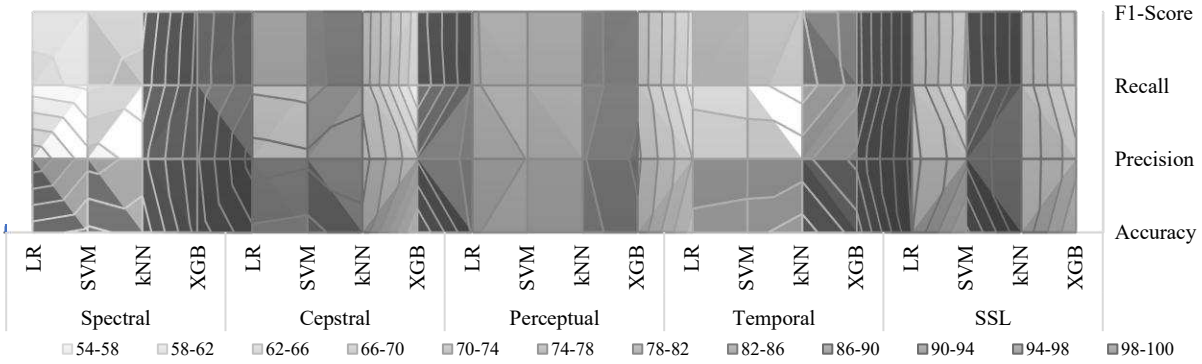


Figure 8: Heatmap visualization of classification metrics across feature domains and classifiers trained on the original dataset and tested on the augmented dataset

3.3. Comparative Analysis

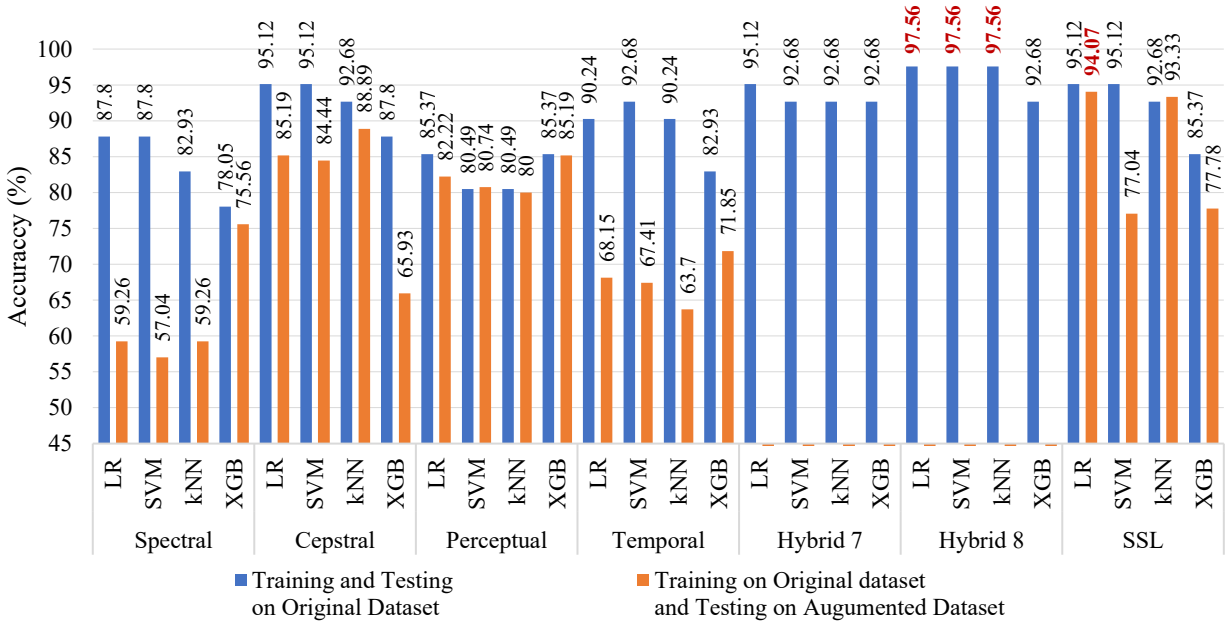


Figure 9: Accuracy comparison of feature groups and classifiers under in-domain and cross-domain evaluation.

Figure 9 presents a comparative analysis of classification performance across different feature representations under both in-domain (training and testing on the original dataset) and cross-domain (training on the original dataset and testing on the augmented dataset) evaluation settings. Blue bars indicate in-domain performance, while orange bars denote performance under cross-domain shift. The results reveal several important trends relevant to robust tabla stroke classification.

First, the effectiveness of feature synergy is clearly evident. Hybrid multidomain feature sets consistently outperform individual feature domains under in-domain settings. In particular, the Hybrid 8 configuration achieves the highest classification accuracy across all classifiers, indicating that the integration of spectral, temporal, cepstral, and perceptual descriptors provides a richer and more discriminative representation of tabla strokes. This confirms that no single feature domain is sufficient to capture the full acoustic complexity of tabla timbre, and that combining complementary cues significantly enhances in-domain performance.

Second, the experiments highlight the robustness of cepstral and SSL representations under domain shift. While spectral and temporal features exhibit substantial degradation when evaluated on augmented data, cepstral features maintain comparatively higher accuracy. This stability reflects the perceptual design of MFCCs, which emphasize spectral envelopes while suppressing fine-grained variability introduced by noise and minor pitch perturbations. SSL representations further improve robustness, achieving the highest cross-domain performance among all feature types. By learning from raw waveform structure using contextual prediction objectives, SSL models capture higher-level spectral-temporal dependencies rather than relying on narrowly defined class-specific cues. Consequently, SSL embeddings tend to suppress nuisance variability such as background noise, minor pitch shifts, and tempo fluctuations, leading to stronger generalization across recording conditions. This suggests that SSL embeddings capture higher-level, domain-invariant acoustic structures that generalize well across recording conditions.

A clear trade-off emerges between in-domain accuracy and cross-domain generalization. While hybrid feature sets achieve near-ceiling performance in matched conditions, their performance declines under

domain mismatch. Conversely, SSL-based representations, although slightly inferior to Hybrid 8 in in-domain accuracy, demonstrate superior generalization to augmented data. This makes SSL representations more suitable for real-world deployment where recording conditions are variable. To further analyze stroke-wise behavior, Figure 10 presents confusion matrices for the best-performing configurations under each evaluation setting. The left matrix corresponds to the Hybrid 8 (SVM) model, which achieved the highest in-domain accuracy (97.56%). The right matrix corresponds to the SSL + LR model, which demonstrated the strongest cross-domain robustness (94.07%).

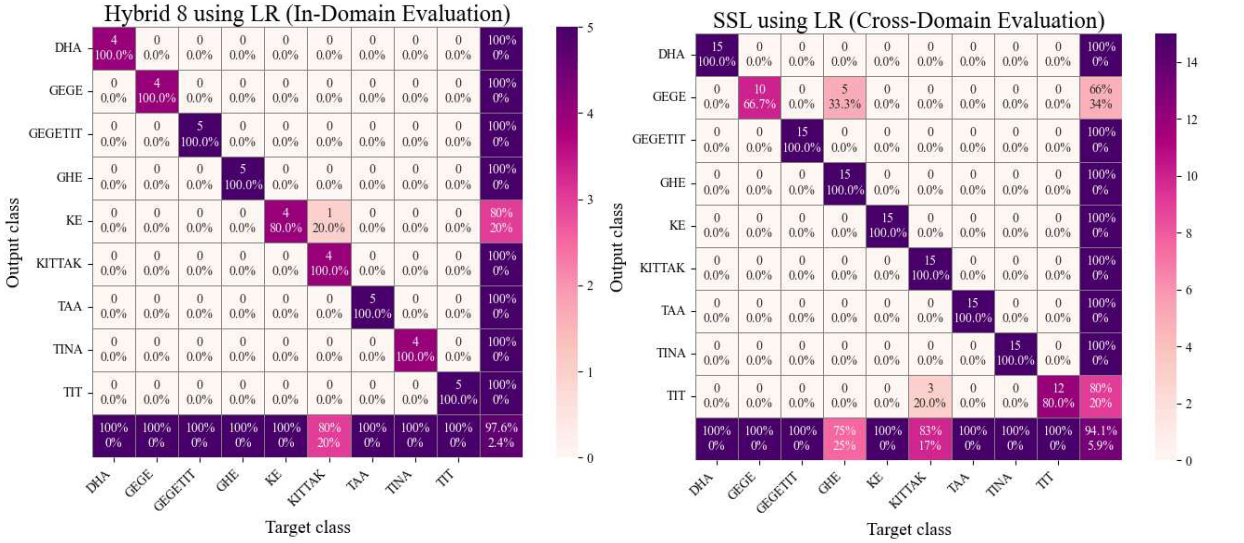


Figure 10: Stroke-wise confusion matrices for the best performing models under in-domain and cross-domain evaluation.

Overall, the comparative analysis demonstrates that multidomain feature integration maximizes discriminative performance under controlled conditions, whereas self-supervised representation learning offers superior resilience to acoustic variability. Balancing these complementary strengths is essential for developing robust and generalizable tabla stroke classification systems.

4. Conclusion

This work presented a comprehensive framework for robust tabla stroke classification under simulated domain shift conditions. Through the integration of multidomain handcrafted features, spanning spectral, temporal, cepstral, and perceptual descriptors together with self-supervised representation learning, we systematically evaluated both in-domain performance and cross-domain robustness. The experimental results demonstrate that multidomain feature synergy significantly enhances in-domain classification performance. The Hybrid 8 configuration achieved near-ceiling performance under matched training and testing conditions, confirming that complementary acoustic cues are essential for capturing the complex timbral characteristics of tabla strokes. However, performance degradation under simulated domain shift highlights the vulnerability of certain handcrafted representations, particularly spectral and temporal features. In contrast, SSL representations exhibited superior robustness when evaluated across domains. Although slightly inferior to Hybrid 8 in in-domain accuracy, SSL representations achieved the highest cross-domain performance, demonstrating their ability to capture domain-invariant spectral-temporal structures from raw waveform inputs. These results underscore the importance of representation learning in mitigating the domain shift problem, which remains a critical challenge in real-world percussion analysis.

The results reveal a clear trade-off between maximum discriminative accuracy and robustness to recording variability. While handcrafted multidomain features maximize accuracy under controlled conditions, self-supervised representations provide stronger generalization in acoustically diverse environments. Future work will focus on expanding the dataset to include multiple instruments and performers, and developing unified architectures that fuse handcrafted and SSL embeddings within end-to-end learning frameworks. Extending the framework toward continuous stroke segmentation and real-time transcription remains an important direction for future work.

5. References

[1]. Ram, S.: A Study of Struggle in Versatility in Tabla Performance. ShodhKosh: Journal of Visual and Performing Arts. 3(1). 218-223. (2022). doi:10.29121/shodhkosh.v3.i1.2022.77

[2]. Patranabis, A., Banerjee, K., Midya, V., Chakraborty, S., Sanyal, S., Banerjee, A., Sengupta, R., Ghosh, D.: Harmonic and Timbre Analysis of Tabla Strokes. arXiv. (2015).

[3]. Gupta, S., Srinivasamurthy, A., Kumar, M., Murthy, H. A., X., S.: Discovery of syllabic percussion patterns in tabla solo recordings. In: Proc. of the 16th Int. Society for Music Information Retrieval Conf.,.385-391. Malaga, Spain. (2015).

[4]. Rohit, MA, Bhattacharjee, A., Rao, P.: Four-way classification of tabla strokes with models adapted from automatic drum transcription. In: Proc. 22nd Int. Soc. Music Inf. Retrieval, pp. 19-26. Online. (2021).

[5]. Sharma, A. K., Aggarwal, G., Bhardwaj, S., Chakrabarti, P., Chakrabarti, T., Abawajy, J. H., Bhattacharyya S., Mishra, R., Das, A., Mahdin, H.: Classification of Indian Classical Music with Time-Series Matching Deep Learning Approach. IEEE Access. 9, 102041-52. (2021). doi: 10.1109/ACCESS.2021.3093911

[6]. Grey J. M.: Multidimensional Perceptual Scaling of Musical Timbre, Journal of the Acoustical Society of America. 61, 1270-77. (1977).

[7]. McAdams, S., Giordano, B. L.: The perception of musical timbre. In Susan Hallam, Ian Cross, and Michael H. Thaut (eds), Oxford Handbook of Music Psychology, Oxford Library of Psychology (2008). doi:10.1093/oxfordhb/9780199298457.013.0007.

[8]. Chordia, P.: Segmentation and recognition of tabla strokes. In: Proc. of the 6th Int. Society for Music Information Retrieval Conf., pp. 107-14. London (2005)

[9]. Gillet, O., Richard, G.: Automatic labelling of tabla signals. In: Proc. of the 4th International Conference on Music Information Retrieval. Baltimore. (2003).

[10]. Deolekar, S., Abraham, S.: Tree-based classification of tabla strokes. Current Science, 115(9), 1724-31. (2018). https://www.jstor.org/stable/26978488

[11]. Gowriprasad, R., Aravind, R., Murthy, H. A.: Structural Segmentation and Labeling of Tabla Solo Performances. arXiv preprint. arXiv:2211.08790. (2022).

[12]. Tzanetakis, G., Cook, P.: MARSYAS: A framework for audio analysis. Organised Sound, 4(3), 169-175 (2000). doi: 10.1017/S1355771800003071.

[13]. Herrera-Boyer, P.: Setting Up an Audio Database for Music Information Retrieval Benchmarking. Technical Report, Music Technology Group, Universitat Pompeu Fabra. (2002).

[14]. Mayya, R., Venkataraman, V., P. R., A., Darapaneni, N.: A novel Bi-LSTM and transformer architecture for generating tabla music. arXiv:2404.05765 (2024).

[15]. Kodag, R. B., Arora, V.: Meta-Learning-Based Percussion Transcription and Tāla Identification from Low-Resource Audio. arXiv preprint. arXiv:2501.04742. (2025).

[16]. Huh, M., Ray, R., Karnei, C.: A comparison of speech data augmentation methods using S3PRL toolkit. arXiv preprint (2023).

[17]. Zhou, Z., Chen, J., Wang, N., Li, L., Wang, D.: Adversarial data augmentation for robust speaker verification. arXiv preprint (2024).

[18]. Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Transactions on Speech and Audio Processing, 9(5), 504–512. (2001). doi: 10.1109/89.928915

[19]. Cohen, I.: Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. IEEE Transactions on Speech and Audio Processing, 11(5), 466–475. (2003). doi: 10.1109/TSA.2003.811544

[20]. López-Espejo, I., Joglekar, A., Peinado, A. M., Jensen, J.: On Speech Pre-emphasis as a Simple and Inexpensive Method to Boost Speech Enhancement. IEEE Open Journal of Signal Processing. 5, 1–14. (2024). doi: 10.1109/OJSP.2024.3352149

[21]. Gourisaria, M. K., Dey, S., Bhowmik, S.: Comparative Analysis of Audio Classification Using MFCC and STFT-Based Features. SN Computer Science. 5(1), 49. (2024). doi: 10.1007/s43926-023-00049-y

[22]. Zhang, X., Yu, D., Wang, Y.: Time-Frequency Representation Learning for Audio Classification Using STFT-Based Neural Front-Ends. Neural Computing and Applications. 36, 15987–16002. (2024). doi: 10.1007/s00521-023-09162-4

[23]. Mahata, K., Saha, G.: Speech and Music Signal Analysis Using Short-Time Spectral Processing Techniques. Circuits, Systems, and Signal Processing. 42, 4893–4912. (2023). doi: 10.1007/s00034-023-02471-2

[24]. Voran, S. D.: Nonuniform Phase Behavior of Short-Time Fourier Transform Coefficients in Audio Signals. IEEE Signal Processing Letters. 31, 180–184. (2024). doi: 10.1109/LSP.2024.3360182

[25]. Zhao, Y., Wang, D., Chen, J.: Deep Learning Based Speech Enhancement Using Short-Time Spectral Amplitude Estimation. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 31, 2210–2223. (2023). doi: 10.1109/TASLP.2023.3279814

[26]. Zeng, Y., Wang, J., Chen, S.: Spectral feature learning for robust audio classification using deep representations. IEEE Transactions on Audio, Speech, and Language Processing, 31, 1123–1135. (2024).

[27]. Liu, X., Li, Q.: Deep spectral aggregation for audio scene analysis. IEEE Transactions on Audio, Speech, and Language Processing, 32, 456–467. (2025).

[28]. Jensen, K., Arnspang, J.: Binary Decision Tree Classification of Musical Sounds. In: Proceedings of the International Computer Music Conference (ICMC), Beijing, China. (1999).

[29]. Zhao, J., Ren, Z.: Advanced cepstral normalization for robust speaker characterization. IEEE Transactions on Audio, Speech, and Language Processing, 32, 725–735. (2025).

[30]. Patel, N., Sharma, M.: Joint cepstral and modulation spectrum features for noisy speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 31, 342–353. (2024).

[31]. Goyal, M., Sahidullah, M.: Improved cepstral parametrization using perceptual filterbanks. IEEE Transactions on Audio, Speech, and Language Processing, 32, 856–867. (2025).

[32]. Kumar, R., Singh, P.: Temporal dynamics modelling with recurrent networks for speech event detection. IEEE Transactions on Audio, Speech, and Language Processing, 33, 589–599. (2025).

[33]. Hsu, W.-N., Zhang, Y.: Temporal structure learning in end-to-end spoken language recognition. IEEE Transactions on Audio, Speech, and Language Processing, 32, 201–214. (2025).

[34]. Yu, T., Wang, D.: Perceptually-weighted spectral features for speech enhancement and intelligibility improvement. IEEE Transactions on Audio, Speech, and Language Processing, 31, 998–1010. (2024).

[35]. Mandal, A., Saha, G.: Psychoacoustic time–frequency feature fusion for emotion recognition in speech. IEEE Transactions on Audio, Speech, and Language Processing, 33, 1877–1889. (2025).

[36]. Scheffé, H.: The Analysis of Variance. John Wiley & Sons. (1959). doi: 10.1002/9780470754099.

[37]. Lowry, R.: Concepts and Applications of Inferential Statistics – Chapter on One-Way ANOVA. Vassar College. (2014). doi: 10.4135/9781412983907.

[38]. Kutner, M. H., Nachtsheim, C. J., Neter, J., Li, W.: Applied Linear Statistical Models. McGraw-Hill Education. (2005). doi: 10.1036/007310874X.

[39]. Schneider, S., Baevski, A., Collobert, R., Auli, M.: wav2vec: Unsupervised Pre-training for Speech Recognition. Interspeech, 3465–3469. (2019). doi: 10.21437/Interspeech.2019-1873.

[40]. Chen, Y.-A., et al.: Data2vec: A General Framework for Self-Supervised Learning in Speech, Vision and Language. International Conference on Machine Learning, 3489–3500. (2022). doi: 10.48550/arXiv.2202.03555.

[41]. Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., et al.: HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3451–3460. (2021). doi: 10.1109/TASLP.2021.3122291.

[42]. Valero, X., Alias, F.: Gammatone Cepstral Coefficients: Biologically Inspired Features for Non-Speech Audio Classification. IEEE Transactions on Multimedia, 14, 1684–1689. (2012). doi: 10.1109/TMM.2012.2224982.

[43]. Palaniappan, R., Sundaraj, K., Arjunan, S. P.: A Comparative Study of the SVM and K-NN Machine Learning Algorithms in Respiratory Sound Classification using MFCC Features. Expert Systems with Applications, 41, 2537–2542. (2014).

[44]. Bianco, M. J., Gerstoft, P., Traer, J., Ozanich, E., Roch, M. A., Gannot, S., Deledalle, C. A.: Machine Learning in Acoustics: Theory and Applications. Journal of the Acoustical Society of America, 146, 3590–3628. (2019).

[45]. Lu, J., et al.: Comparison between Logistic Regression and Machine Learning Algorithms in Noise-Induced Hearing Loss. Frontiers in Public Health, 13, 12048554. (2025).

Funding Declaration

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.