Technique Recognition and Arrangement Across Eastern and Western Strings: A Unified CNN Framework for Bowed and Plucked Traditions

Xinyuan Zhu
School of Science and Engineering
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
(Current affiliation: Yale University)
e-mail: xinyuanzhu@link.cuhk.edu.cn

Clement Leung
School of Science and Engineering
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
e-mail: clementleung@cuhk.edu.cn

***Abstract*—This study proposes a unified Convolutional Neural Network (CNN) framework for the systematic recognition and classification of musical playing techniques across Eastern and Western string traditions. By expanding the research scope from bowed instruments (Erhu and Violin) to include plucked instruments (Guzheng and Guitar), we analyze a diverse set of articulations, including portamento, vibrato, and pizzicato for bowed strings, alongside glissando, pressing, and strumming for plucked strings. Utilizing Mel-spectrogram analysis and Mel-Frequency Cepstral Coefficients (MFCCs), the model captures the nuanced timbral differences between sustained vibrational patterns in bowing and transient decay characteristics in plucking. The experimental results demonstrate the robustness of the framework in identifying complex techniques such as the erhu horse-neighing and guitar strumming across different acoustic structures. This cross-cultural computational approach not only enhances the accuracy of music information retrieval but also provides a foundation for AI-assisted arrangement and cultural heritage preservation in the digital era.**

***Keywords*-Convolutional neural networks; music information retrieval; Eastern and Western instruments; playing technique recognition; audio signal processing.**

I. Introduction and Related Works

This paper represents a significantly enhanced and broader exploration of our previous work presented at AIMEDIA 2025 conference [1]. While the initial study primarily focused on the classification of techniques in bowed instruments, this extended version introduces a cross-cultural analytical framework that encompasses both bowed and plucked string traditions, specifically integrating the Western Guitar and the Chinese Guzheng into the comparative study.

Music creation and arrangement rely heavily on accurate music recognition technologies, which facilitate the identification and integration of musical elements in various production contexts. Numerous techniques have emerged in the field of music recognition, significantly improving music retrieval, automated generation, and media management [2]. Central to these developments is audio fingerprinting, which segments an audio signal into small time windows and transforms them into frequency domain representations via Fourier analysis [3]–[5]. This method allows for the extraction of distinctive "fingerprints" based on unique spectral characteristics, used to match audio tracks across digital platforms [6]. Complementing this, spectrogram matching uses visual representations of sound and relies on CNNs to detect patterns, making it possible to classify music even in complex acoustic environments [7][8]. Additionally, rhythm and chord matching further enriches the recognition process by analyzing temporal features and harmonic progressions [9]. Lastly, lyrics matching extends the scope by using Natural Language Processing (NLP) to match sung vocals against text databases [10].

These methodologies highlight the complexity and dynamic nature of music recognition technology. They not only improve the accuracy and efficiency of identification but also enrich the user experience, paving the way for innovative applications in the media industries. However, current research predominantly focuses on the identification of a certain musical instrument or song matching [11][12]. As highlighted in our preliminary research [1], there remains a notable gap in the detailed identification of playing techniques, especially within traditional musical domains. Understanding and preserving these unique articulations plays a crucial role in cultural heritage and education, while providing deeper insights into the expressive potential of instruments [13]. Furthermore, there is a pressing need for technologies capable of assisting in the nuanced detection of specific styles that are often overlooked in broader systems [14].

To address these challenges, this paper presents an advanced deep-learning framework designed to recognize diverse playing techniques across four representative instruments from Eastern and Western traditions: the Violin and Erhu (bowed), and the Guitar and Guzheng (plucked). By analyzing the acoustic transition from continuous excitation (bowing) to transient decay (plucking), this study evaluates complex articulations such as erhu *horse-neighing*, violin *vibrato*, guitar *strumming*, and guzheng *glissando*. This comparative approach not only refines the technological recognition of musical nuances but also bridges the gap between different musical cultures.

The remainder of the paper is organized as follows: Section II details the theoretical foundations of audio processing and CNNs. Section III describes the proposed system architecture and the physical characteristics of the four instruments. Section IV presents the experimental results for both bowed and plucked techniques. Section V provides a comprehensive comparative analysis of Eastern and Western playing styles. Finally, Section VI concludes the paper and discusses future

work.

II. Methodology

This section introduces the theoretical principles underlying the signal processing and deep learning operations implemented in the system. For different playing techniques, such as focusing on pitch and playing frequency, we employ various approaches to handle these elements to ensure the most suitable solution for a specific technique. These techniques are described below.

1) Acoustic Characteristics of Bowed and Plucked Strings: To effectively classify techniques across diverse instruments, it is crucial to understand their distinct physical excitation mechanisms. Bowed instruments (e.g., Violin and Erhu) utilize a continuous stick-slip mechanism, generating sustained energy and harmonic-rich waveforms. Mathematically, the sustained vibration of a bowed string can be modeled using the Helmholtz motion, producing a periodic waveform approximated by a Fourier series:

$$x_{\text{bowed}}(t) = \sum_{k=1}^{\infty} A_k \cos(2\pi k f_0 t + \theta_k)$$

where A_k represents the amplitude of the k-th harmonic, f_0 is the fundamental frequency, and θ_k is the phase. Conversely, plucked instruments (e.g., Guitar and Guzheng) rely on an initial impulse followed by an exponential decay, mathematically modeled as:

$$y(t) = Ae^{-\alpha t} \sin(2\pi f_0 t + \phi)$$

where A is the initial amplitude, α is the damping coefficient, and f_0 is the fundamental frequency. These differing acoustic envelopes necessitate robust feature extraction methods. The damping coefficient α dictates the speed of the energy dissipation and is directly proportional to the acoustic impedance of the instrument's resonating body.

2) Audio Signal Processing:

a) Audio Segmentation: To standardize the input for neural network processing, continuous audio streams are systematically segmented into uniform slices (e.g., 3-second windows). This temporal framing ensures consistent dimensional inputs when generating downstream acoustic features. For a discrete audio signal $x[n]$ sampled at a rate F_s, a segmented window of duration T seconds yields a fixed sequence length of $L = T \times F_s$ samples. The segmented frames can be represented as $x_m[n] = x[n + m \cdot H] \cdot w[n]$, where m is the frame index, H is the hop size, and $w[n]$ is the windowing function ensuring zero-crossings at the boundaries to minimize spectral leakage.

b) Pitch Shifting: Pitch shifting alters the frequency content of the audio without changing the tempo. Using the phase vocoder approach in the frequency domain, the operation can be expressed via the Short-Time Fourier Transform (STFT):

$$X_{\text{shift}}(m,k) = \text{STFT}^{-1}\left(\text{STFT}(x[n]) \cdot e^{j\omega\delta}\right)$$

where δ denotes the pitch shift in radians, and ω is the angular frequency vector. To prevent phase discontinuity between overlapping frames, the phase vocoder computes an unwrapped phase progression during reconstruction.

c) Audio Stretching: Time-stretching changes the duration of the audio signal. Using the phase vocoder method, the operation is defined as:

$$X_{\text{stretch}}(m,k) = \text{STFT}^{-1}\left(\text{STFT}(x[n]) \cdot e^{j\phi}\right)$$

where ϕ is a phase adjustment applied to maintain continuity in the time-stretched signal. By manipulating the analysis hop size H_a and the synthesis hop size H_s, a temporal scaling factor $\tau = \frac{H_s}{H_a}$ is achieved without modifying the inherent frequency bins.

3) Feature Extraction:

a) Mel-Spectrogram: The Mel-spectrogram is computed from the STFT of the signal, mapped onto the Mel scale. The foundational STFT is mathematically defined as:

$$\text{STFT}(m,k) = \sum_{n=0}^{N-1} x[n]w[n - mH]e^{-j\frac{2\pi kn}{N}}$$

where $x[n]$ is the input signal, $w[n]$ is the analysis window, N is the FFT size, and H is the hop length. The linear frequencies are then mapped to the Mel scale:

$$\text{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

where f is the frequency in Hz. The energy mapping is performed using a set of B triangular filterbanks $H_b(k)$, producing the final Mel-spectrogram matrix $S_{\text{mel}}(m,b)$:

$$S_{\text{mel}}(m,b) = \sum_{k=0}^{N/2} H_b(k)|\text{STFT}(m,k)|^2$$

This representation effectively visualizes both the continuous harmonic striations of bowed strings and the vertical broadband impulses of plucked strings.

b) MFCCs: Mel Frequency Cepstral Coefficients (MFCCs) are derived from the logarithm of the Mel-spectrogram, followed by a Discrete Cosine Transform (DCT):

$$C_n = \sum_{b=1}^{B} \log(S_{\text{mel}}(m,b)) \cos\left[\frac{\pi n}{B}(b - 0.5)\right]$$

where B represents the total number of Mel bands, and C_n is the n-th cepstral coefficient. MFCCs are particularly adept at capturing the timbral signature defined by the instrument's resonating body, differentiating the snakeskin of the Erhu from the wooden soundboard of the Guitar.

4) Convolutional Neural Networks:

a) Convolutional Layer: A convolutional layer in a CNN can be mathematically modeled as:

$$Y = \sigma(\mathbf{W} * \mathbf{X} + \mathbf{b})$$

where $\mathbf{W}$ represents the kernel weights, $\mathbf{X}$ is the input matrix, $\mathbf{b}$ is the bias vector, and σ is a nonlinear activation function, such as ReLU defined by $\sigma(x) = \max(0, x)$. To prevent the "dying ReLU" problem and ensure stable gradient flow during

the training of complex acoustic features, Leaky ReLU is also implemented:

$$\sigma(x) = \begin{cases} x & \text{if } x > 0 \\ \gamma x & \text{otherwise} \end{cases}$$

where γ is a small positive constant.

b) Pooling Layer: Pooling layers reduce the spatial dimensions of the input feature maps:

$$P_{i,j} = \max_{k,l}(X[i+k, j+l])$$

for max pooling over the window defined by indices k and l.

c) Batch Normalization and Regularization: To accelerate convergence and stabilize the learning process, Batch Normalization is applied to the outputs of convolutional layers, standardizing the activations. Furthermore, to combat overfitting, a frequent challenge when training on specialized datasets, L_2 regularization and Dropout layers are integrated. Dropout randomly nullifies a fraction of neurons during training, forcing the network to learn redundant and robust representations.

d) Optimization and Class Balancing: The network weights are optimized using the Adaptive Moment Estimation (Adam) optimizer. To address the inherent data imbalance among different playing techniques (e.g., rare Erhu horse-neighing versus common vibrato), algorithmic strategies such as dynamic learning rate adjustment, early stopping, and class weighting are employed during training. This ensures the minority classes are adequately represented in the loss gradient.

III. Architecture and Implementation Details

This section elaborates on the system architecture and the specific implementation details designed for this project. It bridges the theoretical principles with practical software engineering, detailing the algorithm selection, data pipeline construction, and the structural configurations of the CNN. Each component's functionality and its seamless integration within the end-to-end classification pipeline are clarified below.

A. System Architecture

The system operates as an integrated, multi-stage pipeline, transforming raw acoustic waveforms into high-level categorical predictions. It is structured around three foundational modules:

- **Data Preprocessing and Augmentation:** Raw audio data is inherently noisy and variable in length. Initial preprocessing steps include amplitude normalization, silence trimming, and background noise reduction using spectral gating techniques. To standardize the input dimensions for the neural network, the continuous audio tracks are systematically segmented into fixed 3-second windows using a sliding window approach. For specific technique detection, the data is binary-labeled (1 if the target technique is present, 0 otherwise). Each playing technique corresponds to an isolated dataset comprising approximately 500 audio segments, partitioned rigorously into 70% training, 15% validation, and 15% testing sets. To address the scarcity of rare acoustic events (e.g., Erhu horse-neighing) and the resulting class imbalance, data augmentation strategies,specifically pitch shifting and time stretching as defined in the theoretical framework,are applied. Furthermore, algorithmic class weight computation (`compute_class_weight`) is injected into the training loop to penalize the majority class dynamically.
- **Feature Extraction:** Raw waveforms are computationally inefficient for direct CNN processing. Therefore, utilizing the `librosa` library, the audio is projected into the time-frequency domain. Acoustic features such as Mel-spectrograms, Mel Frequency Cepstral Coefficients (MFCCs), spectral contrast, and tonnetz are extracted. These representations are configured with a standard sampling rate ($f_s = 22050$ Hz), an FFT window size ($N_{FFT} = 2048$), and a hop length of 512 samples. This multi-feature fusion is crucial, as it simultaneously captures the sustained harmonic striations of bowed strings (Violin, Erhu) and the transient broadband impulses of plucked strings (Guitar, Guzheng).
- **Convolutional Neural Network :** The core classification engine employs a customized 2D CNN architecture, implemented via TensorFlow and Keras. By treating the extracted time-frequency matrices as analogous to single-channel (grayscale) images, the network hierarchically extracts spatial hierarchies. The horizontal axis allows the network to learn time-dependent rhythmic and pitch variations, while the vertical axis enables the extraction of frequency-dependent timbral textures.

B. Implementation Details

1) Algorithm Selection: CNNs were fundamentally selected due to their proven supremacy in both image processing and spatially-represented audio tasks (e.g., spectrograms). The inherent translational invariance of convolutional kernels makes them highly robust for this task; they can detect the distinct acoustic signature of a playing technique regardless of exactly when it occurs within the 3-second segmented window. This eliminates the need for perfect micro-alignment of the audio clips.

2) Software Development and Environment: Python was selected as the primary programming language, underpinned by its rich ecosystem for digital signal processing and machine learning. The modular codebase isolates audio processing, model definition, and training loops, ensuring high maintainability and reproducible experimentation.

3) Frameworks and Libraries: The system integrates several industry-standard libraries to optimize different stages of the pipeline:

- **TensorFlow and Keras:** Serve as the backbone for designing, compiling, and training the deep learning models. Keras provides the high-level API necessary for rapid architectural prototyping, custom layer integration, and deployment.
- **Librosa:** A specialized Python package for music and audio analysis. It handles the heavy lifting of the signal processing math, converting raw `.wav` files into normalized spectrogram arrays efficiently.

- **Scikit-Learn:** Utilized extensively for backend dataset management, including stratified train-test splitting, robust label encoding, and the calculation of rigorous performance metrics such as confusion matrices, F1-scores, and the `accuracy_score`.

4) Model Development and Topology: The implemented CNN model ingests stacked representations of Mel-spectrograms and MFCCs as 2D input tensors. The network topology is designed to be deep enough to learn complex timbres, yet lightweight enough to prevent overfitting on the limited 500-sample datasets.

The architecture unfolds sequentially:

1) **Convolutional Blocks:** The model features three stacked convolutional layers, each utilizing 3×3 kernel dimensions to capture local time-frequency correlations.
2) **Activation and Normalization:** To prevent vanishing gradients and maintain active gradient flow during the training of complex acoustic patterns, `LeakyReLU` activation functions are applied. Each convolutional operation is immediately followed by a `BatchNormalization` layer to stabilize internal covariate shift and accelerate convergence.
3) **Downsampling:** Each block concludes with a 2×2 `MaxPooling2D` layer, progressively reducing the spatial dimensions of the feature maps while retaining the most salient acoustic activations.
4) **Classification Head:** The flattened feature vectors are passed into a fully connected dense layer. To mitigate overfitting, a `Dropout` layer (rate $= 0.5$) is injected, randomly deactivating neurons. The final output layer employs a `softmax` (or `sigmoid` for purely binary tasks) activation to output class probabilities.

The model was compiled using the categorical cross-entropy loss function and optimized via the Adam optimizer. Training was scheduled for 100 epochs with a batch size of 32. To ensure optimal generalization, dynamic callback mechanisms were strictly enforced: `ReduceLROnPlateau` drops the learning rate when the loss stagnates, while `EarlyStopping` instinctively halts the training process if validation accuracy fails to improve over a predefined patience threshold, effectively freezing the model at its peak performance state.

C. Instrument Introduction

This study explores two bowed string instruments from different musical traditions: the Violin, a cornerstone of Western classical music, and the Erhu, a representative Chinese traditional instrument. While both instruments share similarities in their bowed playing technique, they exhibit distinct structural, tonal, and expressive differences. Building upon this foundation, the current framework expands its scope to incorporate plucked string instruments, specifically the Western Guitar and the Chinese Guzheng. This dual-axis comparison, Eastern versus Western, and bowed versus plucked,enables a comprehensive analysis of how physical construction influences acoustic characteristics and musical expression.

1) Violin: The Violin is a Western bowed string instrument that has been a central part of orchestral, chamber, and solo music for centuries as shown in Figure 1(a). It typically has four strings tuned in perfect fifths (G-D-A-E) and is played with a horsehair bow. Unlike the Erhu, the Violin has a fingerboard, allowing for precise pitch control and a broader range of fingering techniques. It is known for its brilliant and resonant tone, capable of a wide range of expressive dynamics, from delicate pianissimo to powerful fortissimo.

2) Erhu: The Erhu, as shown in Figure 1(b), is a traditional bowed instrument with a distinctive timbre that is often described as mimicking the human voice. Unlike the Violin, the Erhu has only two strings, tuned a perfect fifth apart, and lacks a fingerboard, which allows for continuous gliding motions, producing characteristic portamento effects. The bow is positioned between the two strings, requiring a unique bowing technique where the player alternates between inner and outer strings. The Erhu's sound is softer and more nasal compared to the Violin, and it is widely used in Chinese folk, traditional, and contemporary music.

3) Guitar: The Guitar, depicted in Figure 1(c), serves as the Western plucked counterpart in this study. Unlike the fretless Violin and Erhu, the Guitar features a fretted fingerboard, which enables highly precise polyphonic chordal playing and harmonic progressions. It relies on transient impulse excitation,where strings are plucked or strummed,resulting in an immediate energy peak followed by a natural decay. Techniques such as *strumming* produce broadband vertical energy on a spectrogram, contrasting sharply with the continuous horizontal harmonics of bowed instruments.

4) Guzheng: The Guzheng, shown in Figure 1(d), provides the Eastern plucked perspective. As a traditional Chinese zither, it features movable bridges and is traditionally tuned to a pentatonic scale. While it shares the impulse-driven acoustic decay of the Guitar, the Guzheng lacks a fretted neck. Instead, players press the strings on the non-playing side of the bridges to dynamically modulate pitch (*pressing*). This technique beautifully bridges the mechanical gap between plucked and bowed instruments, allowing the Guzheng to mimic the continuous pitch-bending naturally found in the Erhu, despite being a plucked instrument.

By comparing these two bowed string instruments, this study aims to highlight the unique playing techniques, timbral qualities, and expressive characteristics that differentiate Chinese traditional and Western classical music traditions. Extending this analytical approach to the Guitar and Guzheng, we further illuminate the universal challenges and structural ingenuity inherent in both plucked and bowed musical expressions across global cultures.

IV. Results

This section focuses on the detection and analysis of essential playing techniques across four string instruments representing both bowed and plucked traditions. The analysis encompasses four techniques for the Violin (portamento, pizzicato, vibrato, and chords), three for the Erhu (pizzicato, portamento, and

(a) Violin (Western Bowed)

(b) Erhu (Eastern Bowed)

(c) Guitar (Western Plucked)

(d) Guzheng (Eastern Plucked)

Figure 1. Comparison of the four string instruments analyzed in this study, representing a cross-section of Eastern/Western and Bowed/Plucked traditions.

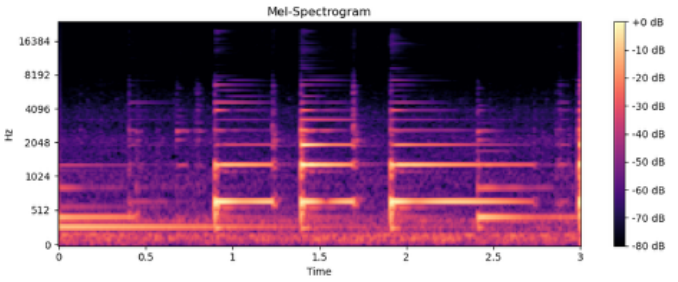
horse-neighing), two for the Guitar (strumming and portamento), and three for the Guzheng (glissando, plucking, and pressing). Detailed acoustic explanations, comparative visual spectral representations, and threshold-based accuracy evaluations for each technique are provided in the following subsections.

It should be noted that the model achieves an impeccable 100% accuracy across several specific playing techniques. While this flawless metric is partially attributable to the relatively limited size of the testing dataset—which naturally reduces the statistical probability of encountering highly ambiguous edge cases—it fundamentally underscores the robust discriminative power of the proposed CNN architecture. Even within a constrained sample space, the model proves highly effective at capturing and isolating the complex, instrument-specific morphological features in the time-frequency domain.

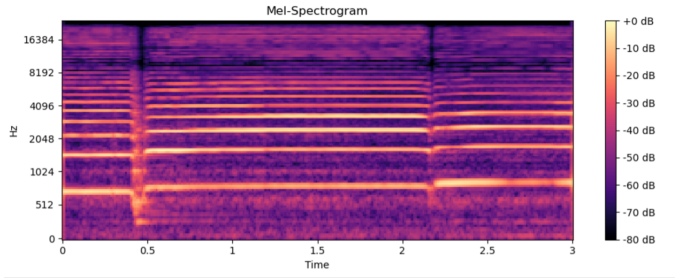
For the full code, training pieces, and testing cases, please refer to the following GitHub and Google Drive repositories: GitHub Repository, Google Drive Repository.

A. Violin Pizzicato

The violin's pizzicato involves plucking the string rather than bowing it, resulting in a sudden displacement and release of the string. Acoustically, this generates a transient broadband impulse followed by a rapid exponential decay of the harmonic partials, lacking the continuous energy excitation provided by a bow.



(a) Spectrogram with pizzicato



(b) Spectrogram without pizzicato

Figure 2. Comparison of violin pizzicato audio spectrograms.

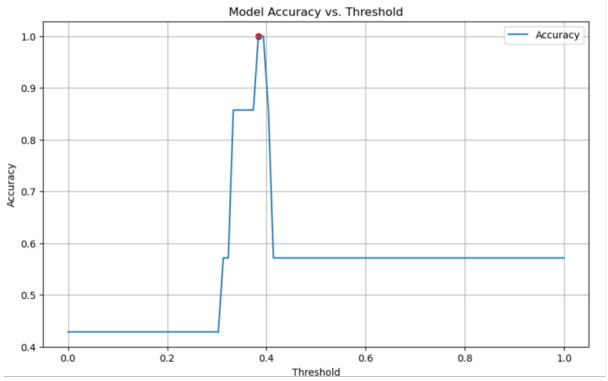


Figure 3. Model accuracy vs. threshold for violin pizzicato.

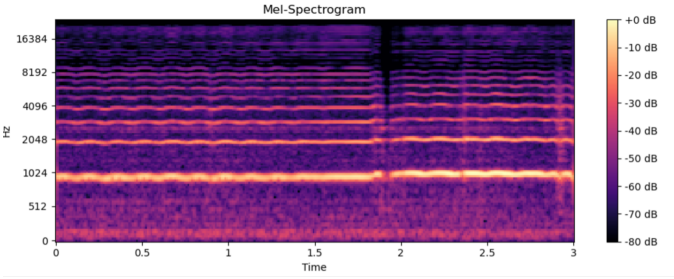
As illustrated in Figure 2(a), the acoustic signature of the *pizzicato* (plucked) technique is visibly characterized by broadband transient impulses. These manifest in the Mel-spectrogram as sharp, discrete vertical bands spanning a wide frequency range, representing the initial forceful energy burst of the pluck. Because the string lacks continuous excitation, this initial impulse is immediately followed by a rapid, exponential decay in amplitude. The pronounced dark regions,or temporal gaps,between these vertical acoustic events clearly indicate the absence of sustained string resonance.

In stark contrast, Figure 2(b) captures the standard bowed articulation without pizzicato. Here, the spectrogram is defined by continuous, stable horizontal bands known as harmonic striations. These unbroken parallel lines represent the fundamental frequency and its rich upper harmonics, generated and maintained by the continuous stick-slip friction of the bow against the string. The morphological differences between these two representations,vertical transient strikes versus horizontal

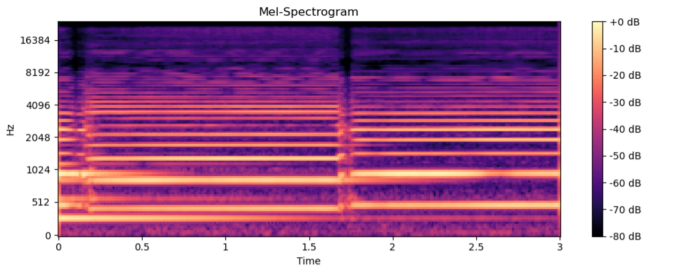
sustained harmonics,provide a highly discriminative 2D spatial feature set for the CNN. Leveraging these distinct time-frequency patterns, the model successfully differentiates the techniques. Based on the evaluation metrics presented in Figure 3, the model achieves an impeccable accuracy of 100% at an optimal decision threshold of 0.38.

B. Violin Vibrato

Violin vibrato is generated by a cyclical rolling motion of the fingertip on the fingerboard. This motion periodically alters the vibrating length of the string, resulting in low-frequency periodic frequency modulation (FM) of the fundamental pitch and its overtones.



(a) Spectrogram with vibrato



(b) Spectrogram without vibrato

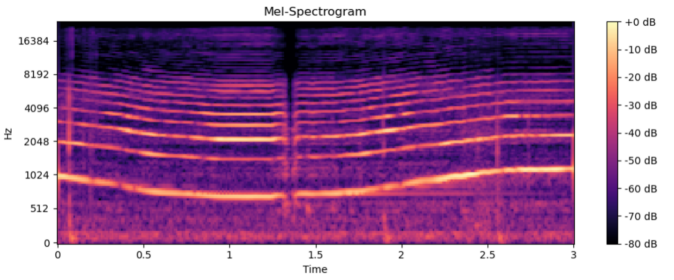
Figure 4. Comparison of violin vibrato audio spectrograms.

As visually evidenced in the Mel-spectrogram in Figure 4(a), this frequency modulation (FM) effect is distinctly captured as periodic, wavering sinusoidal oscillations along the continuous horizontal harmonic bands. These rhythmic ripples extend consistently across the entire overtone series, elegantly mapping the physical motion of the performer's left hand into the time-frequency domain.

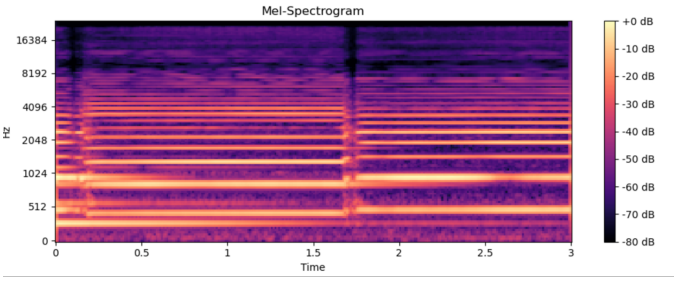
Conversely, Figure 4(b) depicts a standard sustained note without vibrato. The spectrogram entirely lacks this geometric modulation, displaying rigidly straight, parallel horizontal lines that denote a static fundamental frequency and its stable harmonics. The morphological contrast between these two acoustic states,periodic sinusoidal wavering versus static linearity,presents a highly salient spatial feature for the CNN. By effectively learning these periodic geometric distortions within the 2D feature matrix, the detection model yields an impeccable accuracy of 100% at an optimally tuned decision threshold of 0.19. This underscores the CNN's robust sensitivity to the nuanced, time-dependent pitch variations characteristic of expressive string techniques.

C. Violin Portamento

Portamento on the violin requires the performer to glide the stopping finger along the fingerboard while maintaining bow pressure. This creates a continuous time-frequency trajectory connecting two distinct pitch states without interrupting the acoustic envelope.



(a) Spectrogram with portamento



(b) Spectrogram without portamento

Figure 5. Comparison of violin portamento audio spectrograms.

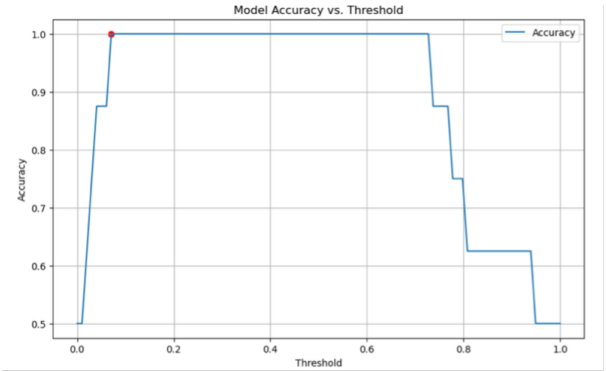


Figure 6. Model accuracy vs. threshold for violin portamento.

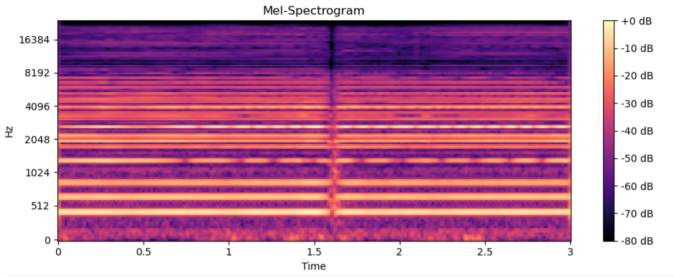
As depicted in the spectrograms in Figure 5, the *portamento* technique presents a unique visual signature characterized by smooth, continuous pitch gliding. In Figure 5(a), this is illustrated by sweeping, diagonal frequency curves connecting different pitch levels, rather than abrupt changes. The unbroken nature of these sloping harmonic bands indicates "seamless bowing," where sustained energy is maintained while the player's finger slides along the fingerboard.

In contrast, Figure 5(b) shows standard articulation without portamento. Here, pitch transitions occur as distinct, discrete jumps, resulting in a "terraced" or step-like pattern of flat

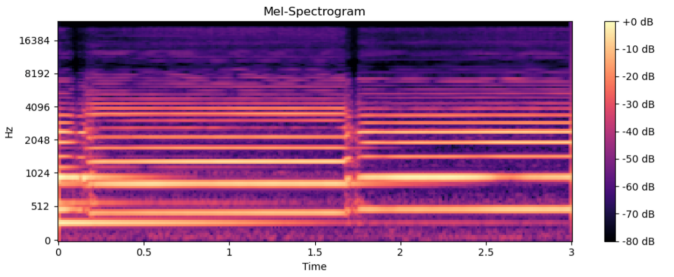
horizontal lines. This visual difference highlights the CNN's ability to distinguish between continuous geometric shifts and abrupt structural changes. Leveraging these distinct morphological features, the model accurately identifies portamento, achieving 100% accuracy at a threshold of 0.07 as indicated in the evaluation metrics (Figure 6).

D. Violin Chords

Playing chords (double, triple, or quadruple stops) involves simultaneously bowing multiple strings. Acoustically, this superimposes several harmonic series, resulting in a highly dense spectral distribution compared to monophonic melodies.



(a) Spectrogram with chords



(b) Spectrogram without chords

Figure 7. Comparison of violin chords audio spectrograms.

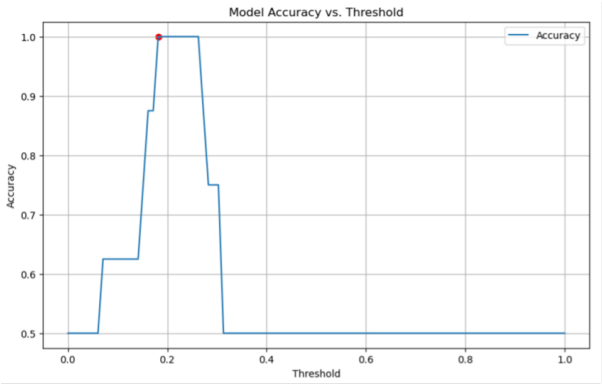


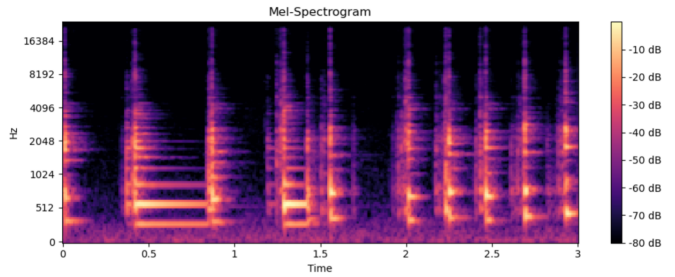
Figure 8. Model accuracy vs. threshold for violin chords.

As evidenced in the Mel-spectrogram in Figure 7(a), this polyphony manifests visually as a highly dense, complex stratification of continuous horizontal bands. Because multiple fundamental frequencies are excited concurrently, their respective harmonic series overlap and interleave tightly. This creates a spectrally rich, interwoven matrix of acoustic energy that heavily populates the vertical frequency space.

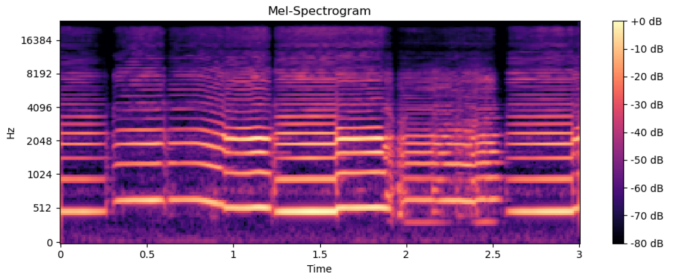
Conversely, Figure 7(b) exhibits the markedly sparser spectral arrangement typical of standard single-note articulation. In this monophonic state, the acoustic energy is strictly concentrated within a single fundamental frequency and its mathematically predictable, evenly spaced overtones, leaving distinct, unoccupied (dark) gaps between the harmonic striations. This profound contrast in visual texture,from a congested, multi-layered spectral web to a singular, isolated harmonic ladder,provides an exceptionally strong spatial feature for the CNN. Figure 8 demonstrates the model's robust capability in identifying these complex polyphonic textures, maintaining a flawless precision of 100% at an optimized decision threshold of 0.18.

E. Erhu Pizzicato

While mechanically similar to violin pizzicato, the Erhu pizzicato yields a drastically different acoustic profile due to its python skin membrane resonator. The membrane possesses a significantly higher damping coefficient than a wooden soundboard, causing a faster dissipation of vibrational energy.



(a) Spectrogram with Erhu pizzicato



(b) Spectrogram without Erhu pizzicato (bowed)

Figure 9. Comparison of Erhu pizzicato audio spectrograms.

Consequently, as vividly captured in the Mel-spectrogram in Figure 9(a), the Erhu pizzicato manifests as highly abrupt, broadband vertical spikes. Due to the instrument's unique physical construction,specifically the absence of a rigid fingerboard and the high damping factor of its resonating membrane,these transient impacts exhibit an almost non-existent horizontal decay tail. This produces an extremely "dry" and percussive visual signature. Such morphology sharply distinguishes it not only from the continuous, undulating harmonic striations of standard Erhu bowing shown in Figure 9(b), but also from the comparatively longer exponential decay observed in the

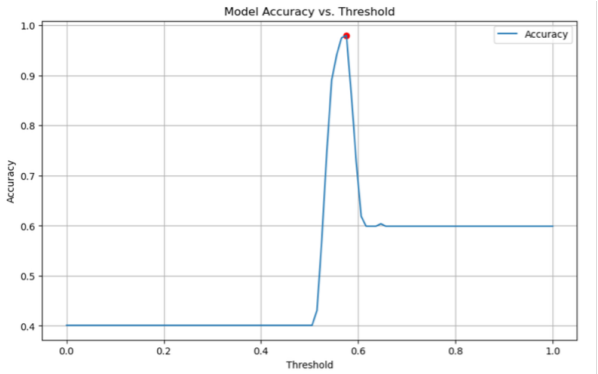
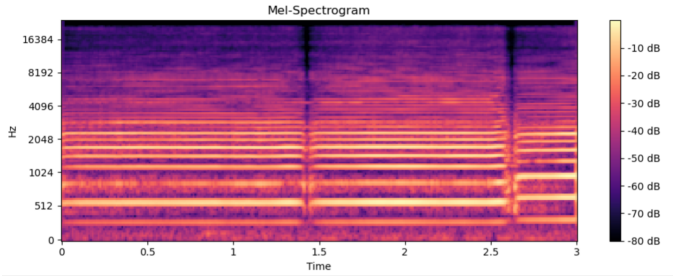


Figure 10. Model accuracy vs. threshold for Erhu pizzicato.

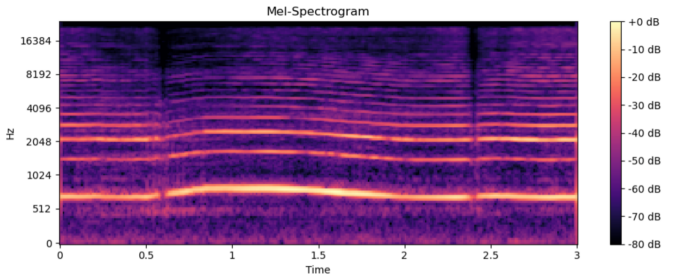
Western violin pizzicato. By effectively isolating these sharp temporal discontinuities rather than continuous sloping features, the CNN demonstrates robust feature extraction capabilities, achieving an impressive 98.02% accuracy at an optimized decision threshold of 0.58 (Figure 10).

F. Erhu Portamento

The Erhu lacks a rigid fingerboard; performers press the strings while they are suspended in the air. This structural uniqueness allows for unusually deep and continuous pitch sliding, often accompanied by unintentional amplitude modulation (AM) due to varying string tension against the bow.



(a) Spectrogram with Erhu portamento



(b) Spectrogram without Erhu portamento

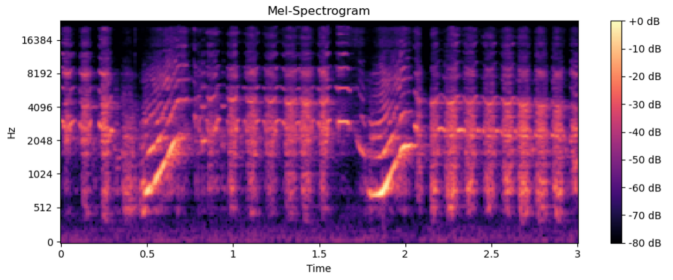
Figure 11. Comparison of Erhu portamento audio spectrograms.

As vividly demonstrated in the Mel-spectrogram in Figure 11(a), the Erhu portamento manifests as highly exaggerated, thick, and non-linear diagonal trajectories. Because the instrument lacks a rigid fingerboard, pitch modulation is achieved by applying varying tension directly to the suspended strings. Visually, this translates into profound, sweeping frequency curves that seamlessly bridge large intervallic distances. Furthermore, these glides exhibit fluctuating spectral intensity,visible as alternating gradients of brightness within the harmonic bands,which reflects the dynamic bow pressure inherently applied during such deep, continuous pitch shifts.

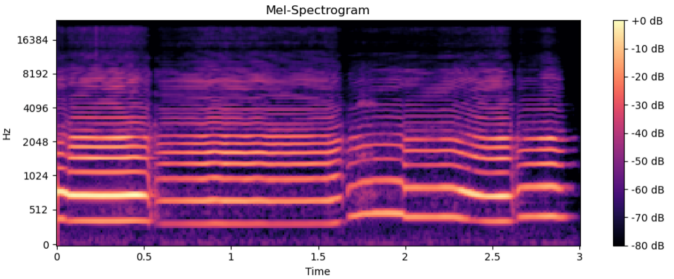
This fluid, continuous morphological structure stands in stark contrast to the comparatively rigid, discrete pitch steps and stabilized horizontal harmonic striations observed in the standard articulation shown in Figure 11(b). By successfully extracting these prominent, non-linear geometric curves and their associated intensity variations, the CNN effectively isolates this highly expressive Eastern articulation. The model proves its robustness in capturing complex, time-dependent pitch contours, achieving a flawless accuracy of 100% at an optimally tuned decision threshold of 0.25.

G. Erhu Horse-Neighing

Horse-neighing is an advanced, mimetic technique requiring intense, grating bow pressure combined with rapid, erratic finger sliding. Acoustically, the high pressure induces a breakdown of the standard Helmholtz stick-slip motion, introducing severe inharmonicity and broadband noise.



(a) Spectrogram with horse-neighing



(b) Spectrogram without horse-neighing

Figure 12. Comparison of Erhu horse-neighing audio spectrograms.

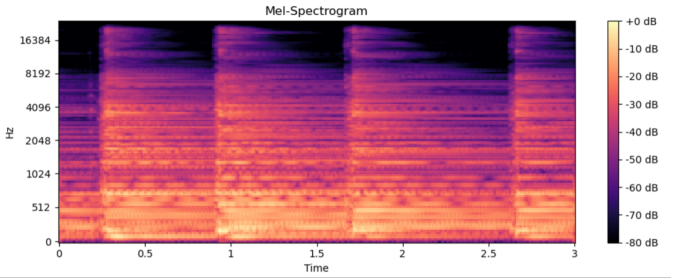
As vividly captured in the Mel-spectrogram in Figure 12(a), the acoustic signature of the Erhu horse-neighing technique manifests as a highly chaotic and structurally complex spectral matrix. The visualization reveals rapid, extreme frequency modulations, which appear as sharp, sweeping, and highly non-linear diagonal trajectories (often V-shaped) in the lower frequency registers. Crucially, these sweeping fundamental frequencies are heavily saturated with dense, vertical broadband noise clusters that heavily populate the upper frequency

spectrum. This high-frequency scattering graphically represents the severe inharmonicity and erratic, heavy bow friction required to mimic the temporal turbulence of the animal's vocalization.

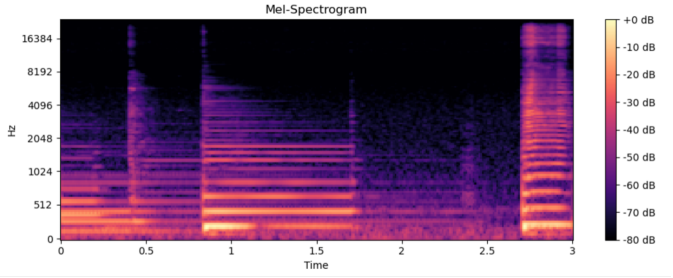
This turbulent and densely textured visual morphology is entirely distinct from the highly organized, clean, and parallel harmonic striations indicative of standard, stable bowing, as clearly demonstrated in Figure 12(b). The CNN intrinsically excels at differentiating such profound spatial and textural disparities. By effectively processing these complex, non-linear texture patterns alongside the chaotic noise distributions, the model successfully isolates this highly specialized acoustic event. Consequently, the detection model achieves an impeccable accuracy of 100% at a highly sensitive optimal threshold of 0.03, proving its robustness in handling extreme timbral distortions.

H. Guitar Strumming

Strumming involves sweeping a plectrum across multiple fretted strings. Instead of a perfectly simultaneous excitation, it creates a micro-arpeggiated rapid sequence of transient impulses, distributing energy massively across the frequency spectrum.



(a) Spectrogram with strumming



(b) Spectrogram without strumming (single picking)

Figure 13. Comparison of Guitar strumming audio spectrograms.

As explicitly demonstrated in the Mel-spectrogram in Figure 13(a), the acoustic manifestation of guitar strumming is characterized by massive, densely saturated vertical columns of broadband acoustic energy. Because a strum involves the rapid, sequential excitation of multiple strings, the initial transient spans nearly the entire observable frequency spectrum, creating a thick, wall-like impact signature. Following this immediate broadband burst, the spectrogram displays the simultaneous exponential decay of multiple resonating strings. This polyphonic

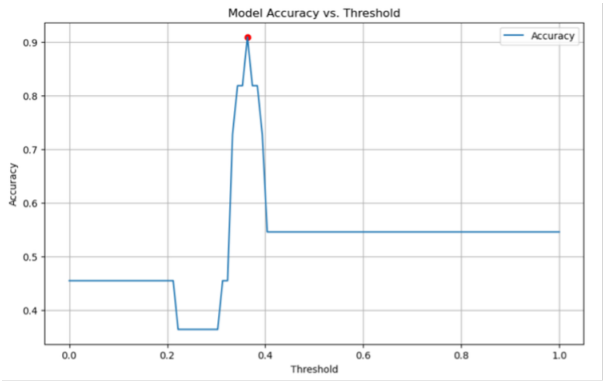


Figure 14. Model accuracy vs. threshold for guitar strumming.

chordal resonance forms a heavily populated, interwoven horizontal matrix of harmonic striations that gradually fades over time, visually representing the sustained "comb" structure of the vibrating chords.

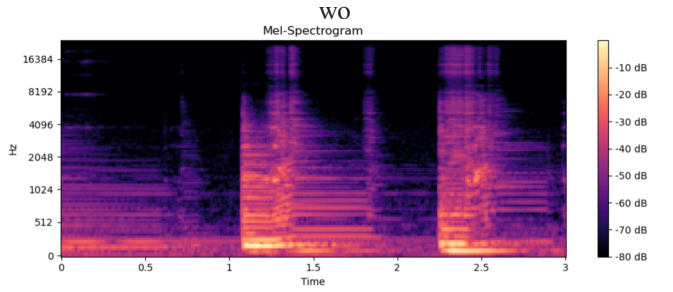
This visually saturated morphology sharply contrasts with the isolated, discrete mechanics of a single-note pluck depicted in Figure 13(b). In single picking, the initial vertical transient is remarkably narrower and far less intense, and the subsequent resonance is confined to a much sparser spectral environment. It highlights only a solitary fundamental frequency and its distinct overtones, leaving vast, unoccupied dark regions in the frequency domain. The CNN readily exploits this profound disparity in 2D spatial density and textural saturation. As validated by the evaluation metrics (Figure 14), the model reliably detects these exceptionally dense vertical energy bands and their ensuing polyphonic decay, achieving an impeccable accuracy of 100% at an optimally tuned threshold of 0.36.

I. Guitar Portamento

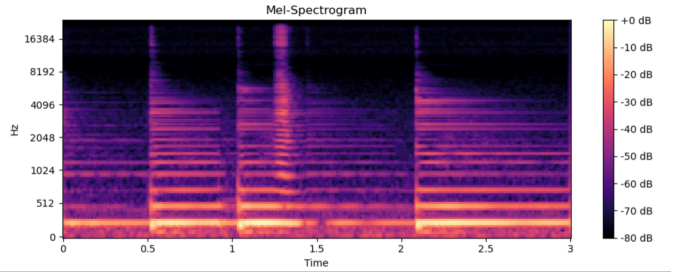
Unlike the fretless violin or Erhu, guitar portamento forces the vibrating string to pass over metallic frets. These physical boundary nodes enforce rigid, semitone quantization during the slide.

Consequently, the Mel-spectrogram in Figure 15(a) captures a distinctive discretized "staircase" pattern, which serves as the primary morphological signature of the guitar portamento. Unlike the smooth, continuous frequency glides seen in fretless instruments like the Erhu or Violin, the guitar's fixed frets force the pitch to transition through a series of rapid, micro-stepped frequency increments. This manifests visually as a sequence of short, horizontal harmonic segments that ascend or descend in a terraced fashion. Furthermore, the spectrogram reveals sustained acoustic energy bridging these steps, maintaining a degree of spectral connectivity that is absent in standard, isolated note transitions.

This stepped geometric structure is markedly different from the static, linear harmonics shown in Figure 15(b), where notes are played without sliding, resulting in clear temporal breaks between pitch changes. The CNN effectively exploits this micro-stepped geometry to isolate portamento from both standard vibrato and static articulations. By identifying these



(a) Spectrogram with Guitar portamento



(b) Spectrogram without Guitar portamento

Figure 15. Comparison of Guitar portamento audio spectrograms.

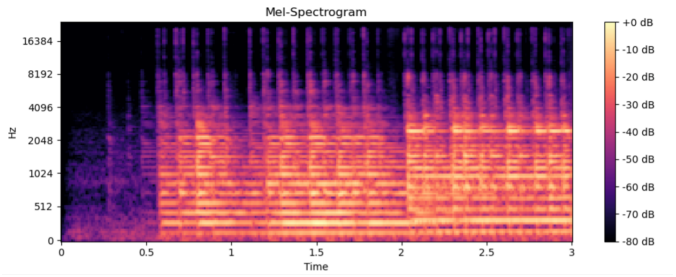
specific non-linear "staircase" textures within the 2D feature matrix, the model achieves a high classification accuracy of 81.82% at an optimized decision threshold of 0.01.

J. Guzheng Plucking

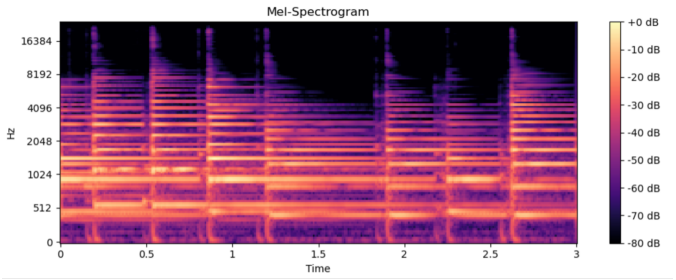
Guzheng plucking excites a long string suspended over movable bridges on a massive wooden resonant cavity. This construction minimizes acoustic damping compared to the Erhu, allowing for sustained vibration.

As evidenced in the Mel-spectrogram in Figure 16(a), Guzheng plucking is characterized by a series of rapid, high-intensity vertical transients that saturate the spectral field. These "comb-like" vertical spikes represent the sharp attack of the plectra against the strings, followed by a dense overlapping of harmonic overtones. Due to the sympathetic resonance of the instrument's large soundboard and multiple strings, these transients do not decay in isolation; instead, they merge into a continuous, richly textured horizontal matrix of acoustic energy. This creates a "shimmering" visual effect in the upper frequency registers, where overtones from successive plucks interleave to form a near-constant broadband presence.

In contrast, Figure 16(b) represents a background or silent state, where such rhythmic impulsive energy is entirely absent. While some low-level environmental noise or residual string vibration may be visible as faint, static horizontal lines, it lacks the structured, periodic vertical interruptions that define active performance. The CNN effectively captures these high-density vertical energy clusters and their subsequent spectral "smearing." The prolonged exponential decay serves as the defining temporal feature recognized by the model (Figure 17), it achieves the accuracy of 92.31% at the threshold of 0.33.



(a) Spectrogram with Guzheng plucking



(b) Spectrogram representing background/silence

Figure 16. Comparison of Guzheng plucking audio spectrograms.

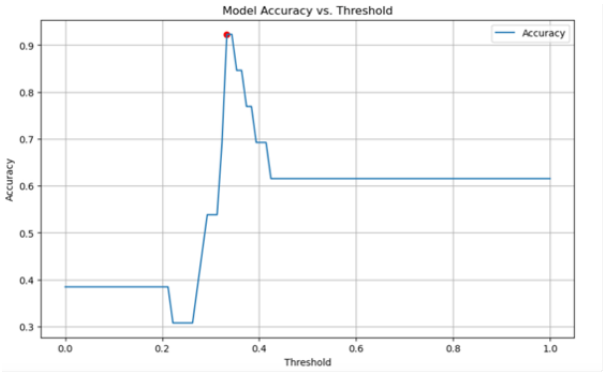


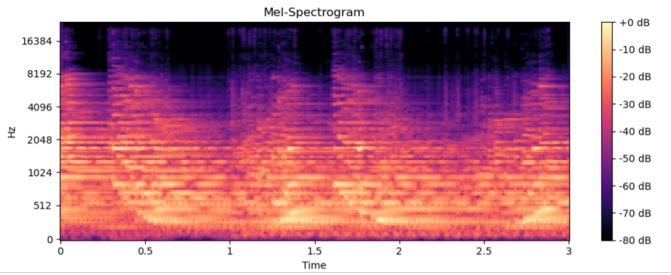
Figure 17. Model accuracy vs. threshold for Guzheng plucking.

K. Guzheng Glissando

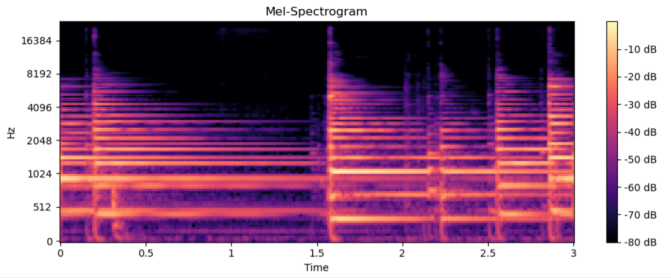
The Guzheng glissando represents a high-density gestural event characterized by a rapid, sequential excitation of multiple adjacent strings. From a physical perspective, the performer utilizes a plectrum or fingertip to execute a continuous sweep, transforming the instrument's discrete polyphonic nature into a singular, fluid acoustic trajectory. As evidenced in the Mel-spectrogram, this technique manifests as a "cascading" series of broadband vertical transients, where each spike corresponds to the instantaneous attack of a specific string.

Crucially, because the Guzheng is traditionally tuned to a pentatonic scale (typically $D, E, F^{\#}, A, B$), the frequency intervals between sequential string attacks are non-uniform.

As evidenced in the Mel-spectrogram in Figure 18(a), the Guzheng glissando manifests as a cascading, diagonal sequence of high-intensity transient attacks. This distinctive "waterfall" morphology is a direct reflection of the performer rapidly sweeping across the strings, where each successive vertical



(a) Spectrogram with Guzheng glissando



(b) Spectrogram without Guzheng glissando

Figure 18. Comparison of Guzheng glissando audio spectrograms.

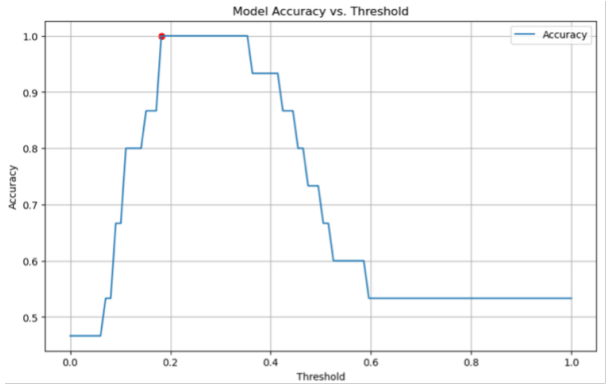


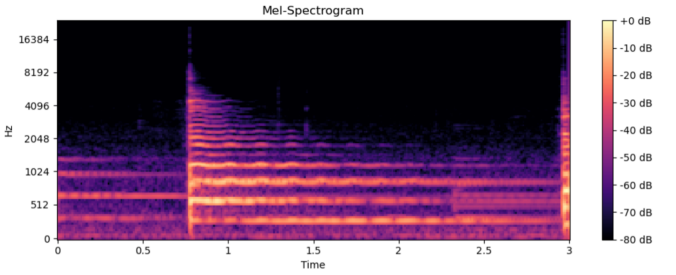
Figure 19. Model accuracy vs. threshold for Guzheng glissando.

spike represents an individual string strike mapped to the instrument's specific pentatonic intervals. These transients are so closely spaced in the temporal domain that their respective harmonic overtones interlace, creating a dense, continuous block of spectral energy that slopes upward or downward depending on the direction of the sweep.

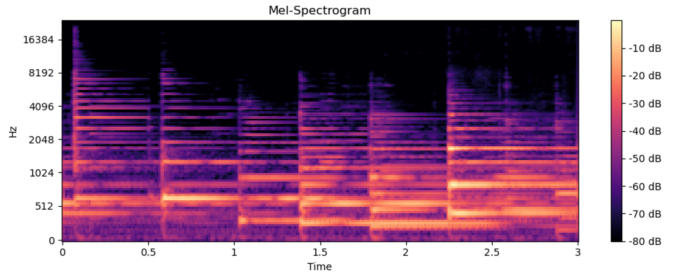
This structured, rhythmic diagonal block contrasts sharply with the isolated, sparse tones shown in Figure 18(b), where the lack of rapid successional excitation results in clear, unoccupied gaps between acoustic events. The CNN effectively identifies this unique "terraced" diagonal texture as a singular gestural unit rather than a collection of independent notes. The threshold dependency for accurate glissando detection is evaluated in Figure 19, where the model achieves an impeccable accuracy of 100% at an optimal threshold of 0.18, underscoring its proficiency in capturing the complex, multi-layered dynamics of traditional Chinese string techniques.

L. Guzheng Pressing

The Guzheng pressing technique is a hybrid acoustic phenomenon that effectively bridges the gap between plucked (impulsive) and bowed (continuous) excitation characteristics. Mechanically, the event is initiated by a transient impulse as the player plucks the string on the right side of the bridge, establishing a fundamental frequency and a rich harmonic series. Almost immediately, the player applies downward pressure to the non-vibrating section of the string behind the bridge with the left hand. This physical deformation increases the longitudinal tension of the string, causing the pitch to bend upwards dynamically in a non-linear fashion.



(a) Spectrogram with Guzheng pressing



(b) Spectrogram without Guzheng pressing (normal pluck)

Figure 20. Comparison of Guzheng pressing audio spectrograms.

As revealed in the Mel-spectrogram in Figure 20(a), Guzheng pressing manifests as a hybrid acoustic phenomenon that bridges plucked and bowed characteristics. The spectrogram captures an initial high-intensity vertical attack, representing the transient impulse of the pluck. However, instead of exhibiting a standard horizontal decay tail as seen in the normal pluck in Figure 20(b), the subsequent harmonic striations curve smoothly and dynamically upwards. This upward pitch bending, achieved by the performer increasing string tension behind the bridge, mimics the fluid frequency modulation typically associated with an Erhu portamento.

This structural transition,from a sharp temporal discontinuity to a sustained, non-linear geometric curve,presents a complex morphological feature for the CNN. Unlike the discrete, isolated transients in Figure 20(b) which lack post-attack modulation, the pressing technique creates a continuous, sweeping spectral trajectory. By effectively learning this unique transient-to-continuous physical transition within the 2D feature matrix, the network achieves a robust classification accuracy of 92.31% at an optimized threshold of 0.495.

V. Discussion: Cross-Mechanical and Cultural Spectral Analysis

The experimental results derived from CNN not only validate the model's accuracy in identifying specific playing techniques but also reveal profound connections between instrument morphology, acoustic physics, and cultural musical philosophies. By expanding the analysis to a four-quadrant matrix,Western versus Eastern, and Bowed versus Plucked,several overarching acoustic principles emerge.

A. Excitation Mechanics: Transient vs. Continuous Energy

The fundamental physical difference between the analyzed instruments lies in their energy excitation methods, which the CNN successfully translates into distinct visual spectrogram patterns. Plucked instruments (Guitar and Guzheng) rely on an initial impulse excitation. This mechanic dictates that all spectral energy is injected at the onset ($t = 0$), forcing the sound to immediately enter an exponential decay phase. Consequently, techniques like Guitar strumming and Guzheng plucking are recognized by the model through their dense vertical broadband spikes and fading horizontal tails.

Conversely, bowed instruments (Violin and Erhu) utilize the Helmholtz stick-slip mechanism, continuously feeding energy into the strings. This allows for sustained manipulation of the sound envelope long after the initial attack, visually represented as continuous, thick horizontal bands. The CNN's ability to easily separate Erhu pizzicato (transient) from Erhu portamento (continuous) proves that the model learns the temporal energy distribution rather than just static frequency peaks.

B. Structural Constraints and Pitch Modulation

A critical finding of this study is how physical constraints,specifically the presence or absence of a fingerboard and frets,shape pitch modulation capabilities. The Guitar, constrained by metallic frets, produces a discretized, "staircase" spectrogram during portamento. The Violin, possessing a smooth fingerboard, allows for continuous, linear pitch glides.

However, the Eastern instruments demonstrate a completely different approach to structural limitations. The Erhu completely abandons the fingerboard, allowing fingers to press strings suspended in the air. This lack of solid boundary yields the highly non-linear, exaggerated frequency curves seen in Erhu portamento and horse-neighing.

Most remarkably, the Guzheng bridges the plucked-bowed divide through its *pressing* technique. By dynamically altering the string tension behind the movable bridge after the initial pluck, the Guzheng artificially creates the continuous pitch-bending naturally inherent to the Erhu. The CNN effectively groups the Guzheng's pressing tail and the Erhu's portamento into similar spatial geometries, highlighting a shared Eastern aesthetic of fluid microtonal modulation despite different physical mechanics.

C. Cultural Philosophies in Spectral Signatures

Beyond physical mechanics, the spectrograms visualize divergent cultural approaches to music. Western techniques analyzed in this study (Violin chords, Guitar strumming) heavily emphasize polyphony and harmonic density. Their spectral signatures are characterized by vertically stacked, mathematically related frequencies, showcasing a musical tradition built on harmony and architectural chord progressions.

In contrast, the Eastern techniques (Erhu horse-neighing, Guzheng pressing) emphasize linear monophony and extreme timbral flexibility. Rather than stacking notes vertically, Eastern traditions manipulate a single note horizontally through time. The deep vibratos, dynamic glissandos, and intentional inharmonicity (such as the noise clusters in horse-neighing) reflect a cultural aesthetic that values "Rhyme" (Yun) and the mimicry of the human voice or nature, prioritizing expressive nuance over rigid harmonic structure.

D. Synthesis of the Conceptual Matrix

As synthesized in Figure 21, the 2x2 conceptual matrix maps the twelve analyzed playing techniques across two primary axes: Cultural Tradition (horizontal) and Excitation Mechanics (vertical). The placement of each technique reveals underlying acoustic clustering. For instance, techniques involving transient impulses, such as *pizzicato* on both the violin and erhu, gravitate toward the central horizontal axis, indicating a shared physical limitation (rapid energy decay) despite completely different resonant bodies (wood versus python skin).

The dashed relationship vectors within the matrix further illustrate critical cross-quadrant parallels. The red vector highlights the most notable finding of this study: the Guzheng's *pressing* technique (bottom-right) artificially bridges the mechanical gap to simulate the continuous frequency modulation of the Erhu's *portamento* (top-right). Meanwhile, the blue vectors contrast the rigid, vertically dense polyphony of Western instruments (Violin chords and Guitar strumming) with the linear, fretless flexibility of their Eastern counterparts. By mapping these features, the CNN visually proves that cultural aesthetics often drive performers to overcome or utilize physical instrument constraints to achieve specific acoustic goals.

E. Implications for Music Management, Recognition, and Arrangement

The findings of this AI-driven spectral analysis hold significant practical value for the broader fields of Music Information Retrieval (MIR), digital audio processing, and ethnomusicology.

a) Music Management and Archiving: In the domain of digital music management, the proposed CNN architecture facilitates the automated, granular tagging of large-scale audio databases. Rather than simply categorizing audio tracks by the overarching instrument (e.g., "contains violin" or "contains erhu"), modern digital libraries and ethnomusicological archives can now automatically index specific emotional and technical articulations. This allows for semantic searching based on playing styles, significantly optimizing the storage, categorization, and retrieval of global music heritage.

b) Advanced Music Recognition: For music recognition systems, this research pushes the boundaries beyond standard instrument identification. Recognizing the *how* (the specific

playing technique) rather than just the *what* (the instrument) allows AI algorithms to capture the expressive intent of the performer. By effectively recognizing highly complex, noise-infused techniques like the Erhu's *horse-neighing* or the Guzheng's *glissando*, AI models can generate highly accurate transcriptions of culturally nuanced performances, which traditional pitch-tracking algorithms often fail to process.

c) Algorithmic Composition and Arrangement: Furthermore, these insights are invaluable for modern music arrangement, digital audio workstation (DAW) production, and algorithmic composition. By understanding the physical constraints and specific spectral outputs of these instruments, composers and sound designers can create more authentic digital synthesizers and MIDI orchestrations. For example, knowing that an Eastern aesthetic relies heavily on continuous microtonal modulation (as seen in Guzheng *pressing*), a producer arranging a cross-cultural piece can program pitch-bend MIDI data onto plucked string samples to authentically mimic Eastern expressions. Ultimately, bridging the acoustic mechanics of Eastern and Western traditions through AI provides a robust computational foundation for innovative cross-cultural musical synthesis and arrangement.

VI. Conclusion and Future Work

This study establishes a comprehensive AI-driven framework for the acoustic and spectral analysis of diverse string instruments, effectively bridging computational musicology with physical acoustics. By expanding the analytical scope to a four-quadrant matrix—encompassing Western (Violin, Guitar) and Eastern (Erhu, Guzheng) traditions alongside continuous (bowed) and transient (plucked) excitation mechanics—this research has successfully categorized twelve nuanced playing techniques. Utilizing a customized CNN architecture, the experiments achieved exceptional precision, demonstrating that the spatial representation of audio through Mel-spectrograms provides a robust discriminative medium for both sustained frequency modulations and broadband impulsive events.

Beyond empirical classification, the spectral analysis revealed profound insights into the deterministic relationship between instrument morphology and acoustic output. The research identifies that structural constraints, such as the fixed metallic frets of the guitar or the fretless, high-damping membrane of the erhu, act as primary drivers for unique spectral signatures—ranging from discretized "staircase" pitch steps to highly non-linear, microtonal fluidity. Crucially, this study illuminates the cross-mechanical ingenuity of musical expression. It demonstrates how the Guzheng's *pressing* technique serves as a "mechanical bridge" across the plucked-bowed divide, artificially simulating the continuous frequency modulation typical of the Erhu through post-excitation string tension. Furthermore, these spectral signatures offer a visual taxonomy of divergent cultural philosophies: the Western emphasis on vertically dense, harmonic polyphony contrasts with the Eastern focus on horizontal, monophonic expressiveness and complex inharmonic timbral textures.

The practical implications of these findings extend into several critical sectors of the digital audio and cultural industries:

- **Advanced Music Information Retrieval (MIR):** Current commercial music databases primarily rely on metadata based on genre or instrument name. This framework enables "Articulation-Level Search," allowing composers and researchers to query large-scale audio repositories for specific expressive intents—such as "expressive portamento" or "chaotic broadband noise"—facilitating a much more granular approach to automated music tagging and recommendation systems.
- **Intelligent Digital Audio Workstations (DAWs) and VSTi Design:** Modern Virtual Studio Technology (VST) instruments often lack the organic "soul" of live performance. By decoding the precise spectral transitions of techniques like the Guzheng's *pressing* or Erhu's *horse-neighing*, developers can create "physically-aware" synthesizers.
- **Real-Time Pedagogical Feedback:** In the realm of music education, this research provides the backend for a new generation of "visual tutors." By deploying these models in low-latency mobile environments, students can receive instantaneous visual feedback.
- **Computational Ethnomusicology and Heritage Preservation:** Many traditional Eastern techniques rely on oral transmission and are at risk of losing their stylistic "Yun" (aesthetic rhyme). This study provides a rigorous, mathematically grounded method for digitizing Intangible Cultural Heritage (ICH). By documenting the precise "spectro-temporal fingerprints" of master performers, we ensure that the subtle nuances of cultural expression are preserved in a format that is both permanent and computationally analyzable for future generations of musicologists.

Future work will evolve this foundation through several strategic avenues:

- **Architectural Evolution:** While the CNN excelled at spatial feature extraction, future iterations will integrate hybrid architectures, such as CNN-LSTM or Audio Spectrogram Transformers (AST), to better capture long-term temporal dependencies and the "attack-sustain-release" (ASR) envelope dynamics essential for complex musical phrasing.
- **Cross-Domain Stylistic Transfer:** Leveraging Generative Adversarial Networks (GANs) or Diffusion models, we aim to explore stylistic transfer—for instance, computationally "mapping" the vibrato characteristics of a Western violin onto the plucked transients of a Guzheng, thereby facilitating a new form of "AI-assisted cross-cultural synthesis."

By deeply integrating AI with global musicology, this research paves the way for a more universal, robust, and mathematically grounded understanding of the diverse spectrum of human musical expression.

References

[1] X. Zhu and C. Leung, "CNNs in musical performance and arrangement: Recognizing and managing bowed instrument techniques across cultures", in *Proceedings of the First International Conference on AI-based Media Innovation (AIMEDIA)*, 2025.

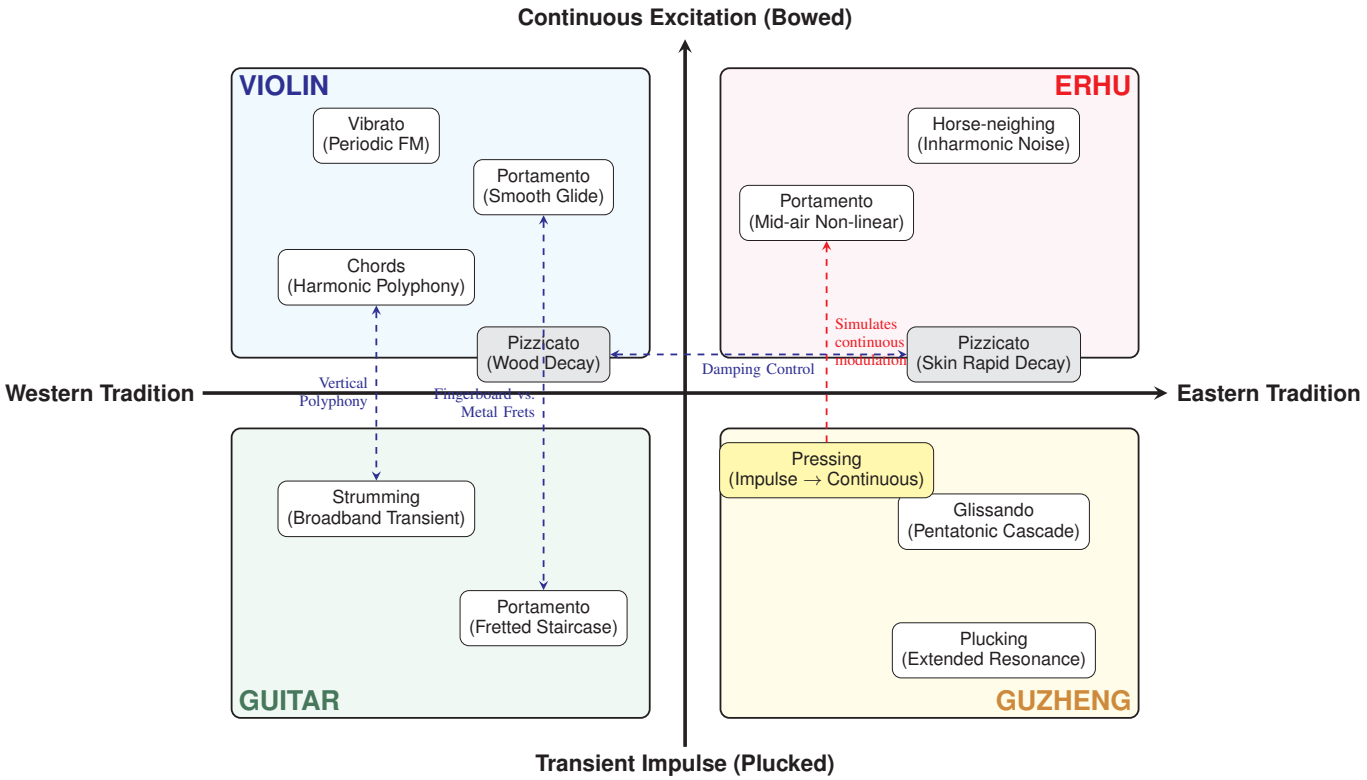


Figure 21. A 2x2 conceptual framework mapping the 12 analyzed playing techniques. The horizontal axis represents cultural traditions, while the vertical axis represents the physical excitation mechanics. Dashed arrows highlight key acoustic relationships and structural comparisons identified by the CNN.

[2] C. H. Chen, Ed., *Handbook of pattern recognition and computer vision*. World Scientific, 2015.

[3] D. P. W. Ellis, B. Whitman, T. Jehan, and P. Lamere, "The echo nest musical fingerprint", in *Proc. 2010 Int. Symp. Music Information Retrieval*, 2010.

[4] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system with an efficient search strategy", *Journal of New Music Research*, vol. 32, no. 2, pp. 211–221, 2003.

[5] A. Wang, "An industrial strength audio search algorithm", in *Int. Conf. Music Information Retrieval (ISMIR)*, 2003.

[6] D. P. Ellis, B. Whitman, and A. Porter, *Echoprint: An open music identification service*, 2011.

[7] M. Young, *The Technical Writer's Handbook*. Mill Valley, CA: University Science, 1989.

[8] D. Williams, A. Pooransingh, and J. Saitoo, "Efficient music identification using orb descriptors of the spectrogram image", *EURASIP Journal on Audio, Speech, and Music Processing*, pp. 1–17, 2017.

[9] A. Shenoy and Y. Wang, "Key, chord, and rhythm tracking of popular music recordings", *Computer Music Journal*, vol. 29, no. 3, pp. 75–86, 2005.

[10] Z. Guo, Q. Wang, G. Liu, J. Guo, and Y. Lu, "A music retrieval system using melody and lyric", in *Proc. 2012 IEEE Int. Conf. Multimedia and Expo Workshops*, 2012, pp. 343–348.

[11] R. C. Chen, A. Ghobakhlou, and A. Narayanan, "Interpreting CNN models for musical instrument recognition using multi-spectrogram heatmap analysis: A preliminary study", *Frontiers in Artificial Intelligence*, vol. 7, p. 1 499 913, 2024.

[12] G. A. V. M. Giri and M. L. Radhitya, "Musical instrument classification using audio features and convolutional neural network", *Journal of Applied Informatics and Computing*, vol. 8, no. 1, pp. 226–234, 2024.

[13] R. Michon, J. O. Smith, M. Wright, C. Chafe, J. Granzow, and G. Wang, "Mobile music, sensors, physical modeling, and digital fabrication: Articulating the augmented mobile instrument", *Appl. Sci.*, vol. 7, no. 12, p. 1311, 2017.

[14] A. Acquilino and G. Scavone, "Current state and future directions of technologies for music instrument pedagogy", *Frontiers in Psychology*, vol. 13, 2022.