Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 18, No. 2, 2026

Multimodal Violin Technique Recognition with Audio and Skeleton Features

Zihang Cao *

School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, Guangdong, 510006, China
*** Corresponding author Email:** 3310529079@qq.com

Abstract: Automatic analysis of violin performance can support music education by helping students and teachers identify playing techniques from ordinary practice videos. This paper focuses on violin technique recognition rather than full automated grading. A simple multimodal framework is used to extract audio descriptors from the soundtrack and skeleton-motion descriptors from video frames, then combine the two modalities for single-label technique classification. The current study considers six dominant technique categories: pizzicato, stopped bowing, plain playing, vibrato, double stops, and spiccato. Experiments on 147 labeled clips show that multimodal fusion achieves 96.60% accuracy and 96.43% macro-F1 under stratified clip-level three-fold cross-validation. A source-video-grouped audit gives 95.92% accuracy and 96.08% macro-F1. The results provide preliminary evidence that audio and skeleton cues are complementary for violin technique recognition.

Keywords: Violin Technique Recognition; Multimodal Learning; Audio-visual Analysis; Skeleton Features; Music Education Technology.

1. Introduction

Technology-assisted music learning increasingly relies on automatic analysis of practice recordings. Existing systems are relatively mature for pitch, rhythm, and transcription-related tasks, but violin technique recognition remains challenging because a technique is expressed through both sound and body motion [1-4]. For example, pizzicato has a clear plucked acoustic attack, spiccato is associated with bouncing bow articulation, vibrato is reflected in periodic left-hand motion and pitch fluctuation, and double stops are mainly defined by simultaneous sounding strings.

This work studies a constrained and measurable problem: classifying short violin performance clips according to the dominant demonstrated technique. It does not attempt to assign a performance-quality score, reproduce an examination rubric, or validate agreement with human grading. This scope is important because technique recognition and automated grading are related but different tasks. The current system is better understood as a recognition component for future educational feedback systems.

2. Related Work

Automatic music performance analysis has often focused on audio-centered tasks such as pitch estimation, rhythm analysis, transcription, genre recognition, and performance assessment [1-4]. These studies show that acoustic descriptors are useful for music understanding, but they do not fully describe instrumental technique because the physical execution of a sound is also important.

Computer vision and pose-estimation methods provide another way to analyze performance motion. MediaPipe and related hand or body pose systems can extract normalized landmarks from ordinary videos [5-8], and recent violin-oriented studies have explored camera-based gesture and posture analysis [9,10]. These works motivate the skeleton stream used in this project.

For violin-specific technique recognition, previous research has studied bowing gesture classification and audio-visual modeling [11-13]. The present project differs by building a small but complete reproducible pipeline that combines audio descriptors, MediaPipe skeleton features, cross-validation, per-class reporting, and explicit discussion of data limitations.

3. Methodology

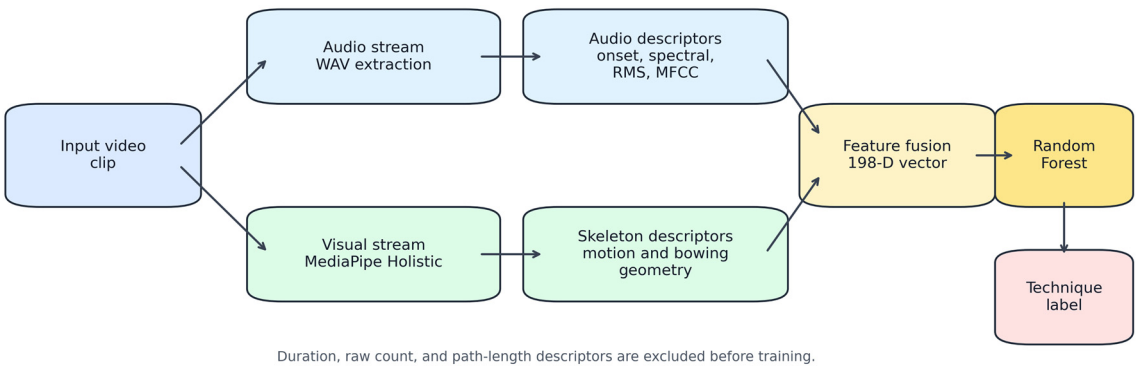


Fig 1. Overview of the proposed audio-skeleton multimodal framework

3.1. Task Definition

Given a short performance clip, the goal is to predict one dominant technique label from six classes: pizzicato (plucked-string articulation; folder boxian), stopped bowing (dungong; short separated bowed notes with clear stop-and-start attacks), plain playing (ordinary bowed playing without the five target special techniques), vibrato (periodic left-hand pitch modulation), double stops (two strings sounded simultaneously), and spiccato (bounced off-string bowing). These classes are not all at the same musical hierarchy, and real violin techniques can overlap. In this study, plain playing clips are retained only when vibrato or another target technique is not the dominant visual or auditory cue; clips with clear sustained vibrato are assigned to vibrato, and ambiguous clips are left out. Tremolo is excluded because only one usable sample is currently available.

3.2. Audio Feature Extraction

Audio is converted to a mono waveform and analyzed with handcrafted signal-processing descriptors. The feature groups include onset rate and onset-envelope statistics, onset-interval statistics, RMS energy statistics, spectral centroid, spectral bandwidth, zero-crossing rate, spectral roll-off, and the mean and standard deviation of the first five MFCC coefficients computed from a mel filterbank. Raw clip duration and raw onset count are excluded from the final classifier because they can reflect recording length rather than technique.

3.3. Skeleton Feature Extraction

Visual features are extracted with MediaPipe Holistic [5-8]. The system tracks upper-body joints and hand landmarks, including both wrists, elbows, shoulders, and five fingertip landmarks for each hand. For each tracked point, it computes coordinate ranges, velocity statistics, acceleration statistics, direction-change statistics, and periodicity. Additional geometric descriptors describe distances and angles related to the right wrist, elbow, shoulder, index finger, and thumb.

All skeleton coordinates are MediaPipe normalized image coordinates rather than raw pixel coordinates. Missing landmark values are represented as NaN during extraction. If a skeleton file contains no usable landmark values, the clip is removed. Otherwise, missing aggregated feature values are filled with zero. Raw landmark counts and path-length features are excluded to reduce frame-count and duration artifacts. Although the coordinates are MediaPipe-normalized, they are not further normalized by instrument scale, performer body size, camera distance, or frame rate.

Table 1. Final feature groups used by the classifier

Modality	Feature group	Dimensions
Audio	Onset, spectral, RMS, roll-off, zero-crossing, MFCC statistics	28
Skeleton	Motion descriptors for upper-body and hand landmarks	160
Skeleton	Bowing-related distances and joint/finger angles	10
Total	Audio-skeleton multimodal vector	198

3.4. Classification

The final classifier is a scikit-learn pipeline [15] consisting of a StandardScaler and a Random Forest classifier with 400 trees, balanced class weights, a minimum leaf size of two, and a fixed random seed [14]. Although Random Forests do not require feature scaling for tree-based splitting, the scaler is kept to make preprocessing explicitly fold-local and to preserve the same pipeline interface for future non-tree baselines. Standardization is fitted inside each training fold and then applied to the corresponding test fold, avoiding preprocessing leakage. The model can output class probabilities, but these probabilities are not treated as calibrated confidence scores.

4. Experiments and Results

4.1. Dataset

The reproducible six-class prepared pool contained 149 WAV candidates. Two stale prepared files were removed because their original source videos could no longer be recovered. After the final re-extraction and validation pass, no additional candidate failed because of missing audio or unusable skeleton features. The final dataset therefore contains 147 usable labeled clips. The clips were collected from publicly accessible violin performance and teaching videos on Bilibili and YouTube, then processed locally to extract WAV audio and MediaPipe skeleton CSV files. Raw WAV durations range from 0.95 to 83.54 s, with a mean of 12.96 s and a median of 7.99 s; audio descriptors are computed from at most the first 20 s of each clip, affecting 20 retained clips. The experiment script writes a data manifest containing the label, prepared audio path, skeleton path, recoverable source-video path, and source group for every retained clip. The source videos and prepared audio files are not redistributed. The current dataset is internal rather than a public benchmark; platform terms, copyright status, performer identity, and piece-level metadata are not yet complete.

Each clip was annotated by the author with one dominant technique label, guided by the source context, visible bowing or fingering motion, and auditory characteristics. The labels have not yet been independently verified by a professional violin teacher or a second annotator, so future work should add multi-annotator agreement analysis before treating the dataset as a formal benchmark.

As an additional internal quality check, a stratified subset of 48 clips was audited with source-video-grouped out-of-fold model suggestions. The model suggestions matched the original labels for 45 of the 48 audited clips (93.75%), and three clips were flagged for manual review. This audit is useful for identifying potentially ambiguous samples, but it is not independent human annotation and does not replace future teacher validation.

Table 2. Dataset composition used in the six-class experiment

Technique	Clips
Pizzicato	27
Stopped bowing	25
Plain playing	33
Vibrato	18
Double stops	23
Spiccato	21
Total	147

4.2. Experimental Setup

Three feature settings are compared: audio-only, skeleton-only, and multimodal fusion. All settings use the same Random Forest configuration and three-fold cross-validation with a fixed random seed. Two split settings are reported. The first is stratified clip-level cross-validation. The second is source-video-grouped cross-validation: all clips with the same recoverable original video path in the manifest are assigned to the same fold. In the cleaned dataset, each retained clip maps to a unique recoverable source-video group, so the grouped audit does not reveal duplicate source-video leakage. However, because verified performer and piece metadata are unavailable, this split should not be interpreted as performer-disjoint or piece-disjoint evaluation and cannot prove cross-performer or cross-piece generalization.

Table 3. Modality ablation results

Split	Modality	Acc. (%)	Bal. Acc. (%)	Macro-F1 (%)
Clip CV	Audio only	91.84 ± 1.67	91.77 ± 1.16	91.66 ± 1.92
Clip CV	Skeleton only	91.84 ± 1.67	91.90 ± 2.37	91.57 ± 2.38
Clip CV	Multimodal	96.60 ± 0.96	96.46 ± 1.45	96.43 ± 1.29
Source-group CV	Audio only	90.48 ± 3.85	89.20 ± 3.99	89.51 ± 4.23
Source-group CV	Skeleton only	91.84 ± 2.89	92.30 ± 2.52	92.28 ± 2.55
Source-group CV	Multimodal	95.92 ± 1.67	96.04 ± 1.88	96.08 ± 1.75

The identical clip-level accuracy values for audio-only and skeleton-only were confirmed from the experiment output. Their balanced accuracy and macro-F1 values differ, so the matching accuracy is a numerical coincidence rather than a copied row. Statistical comparisons are exploratory because only three folds are available. Under source-video-grouped cross-validation, the paired fold-wise macro-F1 gain of multimodal fusion over audio-only is +6.56 percentage points (95% CI: -7.98 to 21.10; paired t-test p=0.192), and the gain over skeleton-only is +3.80 percentage points (95% CI: -5.39 to 12.99; paired t-test p=0.217). Therefore, the results show a positive trend for multimodal fusion but do not justify a strong statistical-significance claim at the 0.05 level. The complementarity claim is therefore based on convergent evidence: non-negative fold-wise differences, different error patterns between modalities, and interpretable feature-importance patterns from both audio and skeleton descriptors.

A cross-modal error analysis gives more concrete support for this interpretation. Under source-video-grouped cross-validation, audio-only and skeleton-only predictions disagreed on 20 of 147 clips. These cases include 10 clips where skeleton-only was correct and audio-only was wrong, 8 clips where audio-only was correct and skeleton-only was wrong, and 2 clips where both single-modality models were wrong but predicted different labels. Among the 20 disagreement cases, the multimodal model correctly classified 16 clips. It preserved the correct label in 10 of the 10 skeleton-correct cases and in 5 of the 8 audio-correct cases; among the remaining 2 both-wrong cases, fusion recovered 1 clip.

Table 4. Per-class multimodal performance under source-video-grouped cross-validation

Technique	Precision	Recall	F1	Support
Double stops	0.958	1.000	0.979	23
Stopped bowing	0.960	0.960	0.960	25
Pizzicato	1.000	0.963	0.981	27
Plain playing	0.969	0.939	0.954	33
Spiccato	0.870	0.952	0.909	21
Vibrato	1.000	0.944	0.971	18

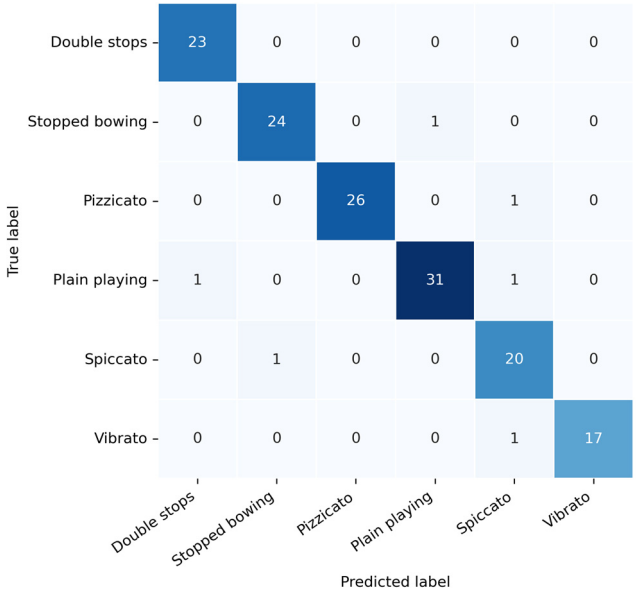


Fig 2. Confusion matrix under source-video-grouped cross-validation

The multimodal model performs best in both split settings. Under source-video-grouped cross-validation, it reaches 95.92% accuracy and 96.08% macro-F1, higher than the single-modality baselines. The feature-importance analysis suggests that spectral-centroid variation, MFCC statistics, onset-envelope variation, right-hand direction changes, wrist velocity, and left-hand periodicity are informative. These observations are consistent with the musical intuition that violin techniques are reflected in both acoustic outcomes and physical execution patterns.

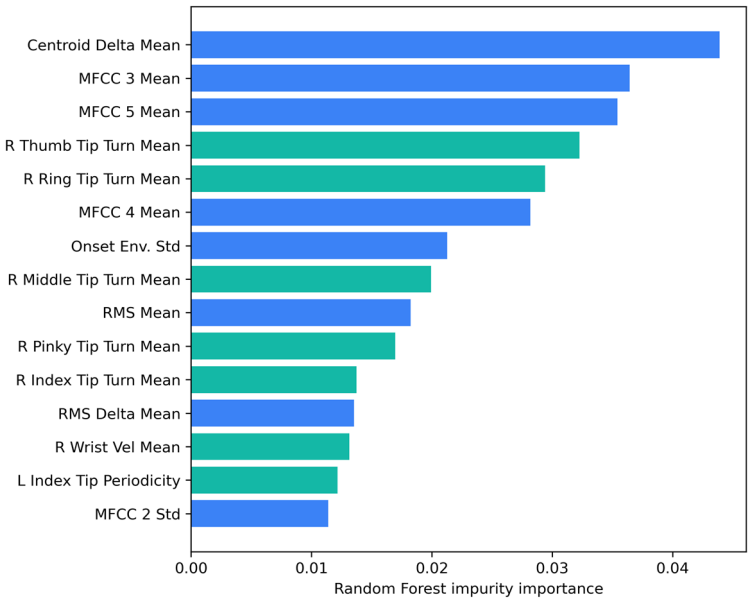


Fig 3. Top feature importances of the multimodal classifier

5. Discussion

The results support the usefulness of combining audio and skeleton features for violin technique recognition. Audio features are strong for techniques with clear acoustic signatures, while skeleton features are helpful for techniques expressed through hand or bowing movement. Although the fold-wise gain is not statistically significant at the 0.05 level with only three folds, the non-negative fold-wise trend, cross-modal disagreement analysis, and interpretable feature patterns provide convergent evidence for complementarity. The 48-clip model-assisted audit also matched 45 original labels, suggesting that the main label taxonomy is mostly coherent while still leaving ambiguous samples for future teacher review. Double stops and pizzicato obtain the strongest F1 scores because they have distinctive cues: double stops are associated with simultaneous-string spectral patterns, while pizzicato produces sharp plucked attacks and motion patterns that differ from ordinary bowed playing. Spiccato has the lowest F1 among the six classes, likely because short bounced-bow passages can share rapid onset patterns with stopped bowing and articulated plain playing. MediaPipe skeleton landmarks also do not directly observe bow hair contact, bow bounce height, or string interaction. The removal of duration, raw count, and path-length features reduces the risk that the classifier relies on trivial recording-length cues.

Several limitations remain. First, the dataset is still small and class-imbalanced, and the current source-group split is based on recoverable source-video paths rather than verified performer or piece identities. The raw material also comes from online platform videos, so copyright, creator permission, and privacy status are not complete enough for public raw-media release. Second, the task is single-label, although real violin techniques can overlap. Third, the labels are dominant-technique annotations rather than independently verified teacher rubrics. Fourth, Random Forest probabilities are not calibrated confidence scores. Finally, skeleton coordinates are MediaPipe-normalized image coordinates, but they are not further normalized by instrument scale, performer body size, camera geometry, or frame rate, which may affect cross-performer and cross-camera generalization.

6. Conclusion

This paper presented a simple multimodal framework for six-class violin technique recognition from performance videos. The proposed system combines audio descriptors with MediaPipe-based skeleton-motion features and uses a Random Forest classifier for interpretable within-dataset classification. On 147 labeled clips, multimodal fusion achieved 96.60% accuracy and 96.43% macro-F1 under stratified clip-level cross-validation, and 95.92% accuracy and 96.08% macro-F1 under source-video-grouped cross-validation. These results suggest that audio-skeleton fusion is a useful candidate component for violin learning support, while full automated grading remains a separate future task requiring teacher rubrics, quality scores, performer- and piece-grouped evaluation, and stronger annotation validation.

References

[1] Zhang, H., Cheung, V. K. M., Nishioka, H., Dixon, S., & Furuya, S. (2025). LLaQo: Towards a query-based coach in expressive music performance assessment. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[2] Schedl, M., Gomez, E., & Urbano, J. (2014). Music information retrieval: Recent developments and applications. Foundations and Trends in Information Retrieval, 8(2–3), 127–261.

[3] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10(5), 293–302.

[4] Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., & Klapuri, A. (2013). Automatic music transcription: Challenges and future directions. Journal of Intelligent Information Systems, 41(3), 407–434.

[5] Lugaresi, C., et al. (2019). MediaPipe: A framework for building perception pipelines. CoRR, abs/1906.08172.

[6] Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.-L., & Grundmann, M. (2020). MediaPipe Hands: On-device real-time hand tracking. CoRR, abs/2006.10214.

[7] Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., & Grundmann, M. (2020). BlazePose: On-device real-time body pose tracking. CoRR, abs/2006.10204.

[8] Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of the CVPR (pp. 7291–7299).

[9] Yoo, S. J., Shrestha, S., Muresanu, I., & Fermuller, C. (2025). VioPose: Violin performance 4D pose estimation by hierarchical audiovisual inference. In Proceedings of the WACV.

[10] Lucena, R., & Ramirez, R. (2024). A machine learning approach to gesture detection in violin performance. In Proceedings of Audio Mostly.

[11] Dalmazzo, D., & Ramirez, R. (2019). Bowing gestures classification in violin performance: A machine learning approach. Frontiers in Psychology, 10, 344.

[12] Wilson, W., Granieri, N., & Ali-MacLachlan, I. (2023). Ubiquitous multimodality as a tool in violin performance classification. In Proceedings of the International Symposium on the Internet of Sounds.

[13] Muhammed, Z., et al. (2024). Ensemble of multimodal deep learning models for violin bowing techniques classification. Journal of Advances in Information Technology, 15(1), 40–48.

[14] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[15] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.