TISMIR

Smartwatch-Based Audio–Gestural Insights in Violin Bow Stroke Analyses

RESEARCH ARTICLE

ubiquity press

WILLIAM WILSON
NICCOLÒ GRANIERI
SAMUEL SMITH
CARLO HARVEY
ISLAH ALI-MACLACHLAN

*Author affiliations can be found in the back matter of this article

ABSTRACT

Following the exposition of quantitative, identifiable idiosyncrasy in violin performance – via neural network classification – we demonstrate that smartwatch-based synchronous audio-gesture logging facilitates interpretable practice feedback in violin performance. The novelty of our approach is twofold: we exploit convenient multimodal data capture using a consumer smartwatch, recording wrist-movement and audio data in parallel. Further, we prioritise the delivery of performance insights at their most interpretable, quantifying tonal and temporal performance trends. Using such accessible hardware to observe meaningful, approachable performance insights, the feasibility of our approach is maximised for use in real-world teaching and learning environments. Presented analyses draw upon a primary dataset compiled from nine violinists executing defined performance exercises. Recordings segmented via note onset detection are subject to subsequent analyses. Trends identified include a cross-participant tendency to 'rush' up-bows versus down-bows, along with lesser temporal and tonal consistency when bowing *Spiccato* versus *Legato*.

CORRESPONDING AUTHOR:
William Wilson
Acoustics and Audio Analysis research lab, College of Computing, Birmingham City University, Birmingham, B4 7RQ, UK
william.wilson@bcu.ac.uk

KEYWORDS:
datasets, neural networks, gestural analysis, computational musicology, violin, IMU sensors, technology-enhanced learning

TO CITE THIS ARTICLE:
Wilson, W., Granieri, N., Smith, S., Harvey, C., & Ali-MacLachlan, I. (2025). Smartwatch-Based Audio–Gestural Insights in Violin Bow Stroke Analyses. *Transactions of the International Society for Music Information Retrieval*, *8*(1), 283–299. DOI: https://doi.org/10.5334/tismir.216

1 INTRODUCTION

The role of motion in violin performance has proved a popular field of study; Palac (1992) attributes this to the visibility of the sound-producing motion – the bow-stroke. Prior works demonstrate the utility of computational technologies for quantitative analyses of the violinist's execution. Early works employed pioneering technical apparatus, including a vibration microscope for observation of the bowed string (Helmholtz, 1895). More recently, works quantifying gestures have employed optoelectronic technologies (Volpe et al., 2017), electromyographic interfaces (Sarasúa et al., 2017) and inertial measurement unit (IMU) sensors (Dalmazzo et al., 2018). Many of these works propose applications in technology-enhanced learning (TEL). Resultant research products remain largely confined to laboratory settings, possibly due to an apparent theme of such works: the dependence upon specialist, niche or otherwise inaccessible devices. To develop a TEL tool for use in teaching and learning environments, we investigate the utility of a mainstream consumer device – the smartwatch – while conducting violin performance analyses. Figure 1 depicts an example multimodal recording, encompassing three-axis gyroscopic data, alongside concurrent audio. Presented analyses indicate quantitative performer idiosyncrasy, via multi-input deep neural network (MI-DNN) classification. MI-DNNs facilitate an assessment of multimodal learning, versus conventional unimodal counterparts (Audio, IMU). We find multimodality does not improve participant recognition accuracy invariably.

Further, we delineate shared and individual performer tendencies via statistical means. Statistically significant trends are identified, including cross-participant tendencies to 'shorten' up-bows versus down-bows, along with greater tonal and temporal variance when bowing *Spiccato* versus *Legato*.

This paper is structured as follows: Section 2 presents a review of literature deemed foundational to our primary analyses. Section 3 recounts primary data capture, detailing our recording and segmentation of concurrent audio and IMU data. Section 4 details processing – or lack thereof – necessitated by each datatype, ensuring suitability for subsequent classification. Section 5 details a comparison of neural architectures trained upon primary data towards participant recognition. Networks include single-input classifiers trained upon lone datatypes (Audio, IMU) and MI-DNNs trained upon both simultaneously. Section 6 details findings of subsequent statistical analyses, with an eye towards the identification of both tonal and temporal performer tendencies, and note-by-note analyses. Section 7 offers a final reflection, detailing a consideration of feasibility.

2 BACKGROUND AND RELATED WORKS

This section provides an overview of concepts foundational to forthcoming analyses, including the production of tone via performer–instrument interaction, audio-gestural approaches towards analysis of this, audio feature–extraction techniques and neural network architectures used in primary classification analyses.

2.1 VIOLIN PERFORMANCE

Most fundamentally, violin performances include series of audible notes – each note being a product of the violinist's bow-stroke. When the bow is drawn across the string, friction causes the string to vibrate audibly.

$$f = \frac{1}{2L}\sqrt{\frac{T}{\mu}} \tag{1}$$

Equation 1: String resonance equation, where f (Hz) is the vibrational frequency, L (m) is the length of the string, T (N) is the tension of the string and μ (kg/m^3) is the mass per unit length of the string (Grimes, 2014).

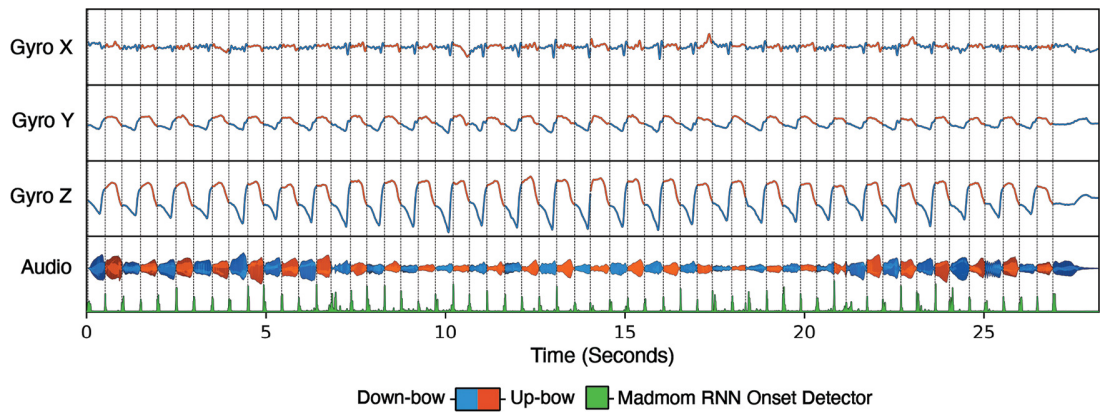

Figure 1 A Synchronous Inertial Measurement Unit–Audio dataset recording, with vertical lines denoting note onsets peak-picked from an onset detection function calculated through use of the Madmom[1] audio signal–processing library.

Audible vibrations of the string are amplified by the resonant instrument body. The frequency of the note produced is a function of the tension, mass and length of the bowed string (Equation 1). By pinching a string against the fingerboard, the violinist effectively shortens the string, producing a higher frequency note than when the string is bowed 'open'; the violinist deliberately – although subconsciously – manipulates parameters/variables of Equation 1, towards an intended musical output. The violinist can further manipulate the timbre of the sound produced by varying gestural aspects of their bow-stroke execution. These include bow speed, pressure and distance from the bridge (Guettler, 2010). Through a precise combination of such gestural parameters, a performer may produce a note embodying one of many established bow articulations. Of these, Perez Carrillo (2013) identifies two categories: on-string and off-string notes.

As an example of off-string notes, Shan et al. (2007) describe *Spiccato* notes as a kind of 'momentary bouncing' of the bow against the string. While the audible note is solely a product of contact with the string, the authors consider *Spiccato* bow-strokes to be composed additionally of 'airborne phases', wherein the bow either approaches or has departed the string. However, the authors note, such phases may prove absent at higher tempi, given uninterrupted contact. Quantifying bowing kinematics at tempi ranging between 144 and 576 bows per minute, the authors note decreased movement extents at higher tempi, along with non-linearities in bow-speed and acceleration across tempo conditions.

Perez Carrillo (2013) defines on-string notes as exhibiting contrasting, sustained contact between the bow and string for the duration of the audible note, citing *Legato* as an example. Quantitavist Guettler (1997) considers consistency in the amount of force applied to the bow a crucial aspect of *Legato* bow-strokes, contributing to characteristic fluency in audible phrasing. In a perhaps contradictory prior description of the *Legato* stroke, Dippel (1987) described these as comprising 'masked bow changes', achieved through a reduction in weight applied to the bow during the transition between notes.

Investigating the influence of gestural execution towards the perception of violin performance, Bugaj et al. (2019) reported higher subjective assessments of 'musicality' amongst participant-evaluators observing performers recorded playing excerpts in a 'high-movement' condition. Participant-evaluators provided no visual stimuli instead reported higher ratings of 'musicality' when listening to 'low-movement' excerpts. Deducing 'implications for practice', the authors caution that consideration of gestural execution may occasionally be neglected in teaching and learning environments, concluding: 'Music educators may wish to stress to students that movement is part of music, and lead students to interpret music through movement[...]' (Bugaj et al., 2019).

2.2 AUDIO-GESTURES

Inherently, audio-gestural approaches towards violin performance analyses necessitate multimodal approaches – typically involving a microphone, alongside one of various gestural sensing technologies. Used concurrently, these facilitate quantitative observation of both performer movements and their resultant tone. Applications range from creative, compositional tools (Thorn, 2018) to analytical tools seeking to quantify aspects of performance – typically, towards pedagogical utility.

Assessing prior kinematic studies of music performance, Shan et al. (2007) described a 'dearth' of quantitative perspectives. While acknowledging the necessity of qualitative approaches, the authors characterise quantitative approaches as objective, offering resultant utility for motor-learning in violin practice. As justification, the authors cite potential applications in injury prevention and the avoidance of discord between subjective teacher/student reflections. Perhaps ironically, Palac (1992) demonstrates the potential for discordance between such quantitative study and established violin pedagogy – deeming elbow positioning idealised by Suzuki 'incompatible' with kinematic criteria, devised by Helmholtz, for the production of 'good tone'.

Towards an understanding of bow-stroke kinematics, Rasamimanana et al. (2007) employed a combination of optical and pressure sensors. One violinist and a violist were recorded performing repeated, successive bow-strokes escalating in tempo before subsequently de-escalating. The authors highlighted biomechanical constraints towards performance gesture execution, quantifying decreased stroke extent at higher tempi – echoing the findings of Shan et al. (2007). Further, the authors characterised bow velocity at higher tempi as approximately sinusoidal; lower tempi bow velocities more closely approximated a square wave, suggesting greater gestural continuity at higher tempi. Despite a gradual transition in tempo, a non-linear fit indicated 'the transition from one [state] to the other is abrupt, with a hysteresis effect'.

In a study of nine expert violinists, Chander et al. (2022) employed a 12-camera marker-based opto-electronic system, observing the influence of musical cadence on gestural expression and temporal execution. Detailing practicalities, the authors disclose a requirement for participants to wear 'dark, close-fitting clothing' to which the researchers affixed reflective markers, alongside their instruments. During calibration, participants were required to assume a reference pose. Participants were instructed to recite excerpts as they would during a solo recital. The authors quantify increased deviation from an established tempo during excerpts incorporating musical cadences, attributing an observed prevalence of *ritardando* (wherein notes are 'stretched' longer

than their notated duration) towards intended musical expression, serving to emphasise cadential content. The authors also report a greater degree of gestural expression during excerpts containing 'global' versus 'local' cadences, quantified via principal component analysis of 'non-technical motion' —a term denoting movement non-crucial for the production of tone.

Leveraging machine learning techniques in TEL applications, prior works have most frequently demonstrated the utility of gestural sensor apparatus towards classification tasks. Sarasúa et al. (2017) and Dalmazzo et al. (2018) employed the Myo armband (Thalmic Labs, Waterloo, ON, Canada), assessing the device's suitability as a data-capture interface towards bow-articulation classification. Dalmazzo et al. (2018) assessed device utility through comparison with a benchmark: an infrared optoelectronic system containing eight cameras. The authors employed a J48 decision tree algorithm, reporting marginally higher classification accuracy with the use of Myo-recorded data (99.847%) versus the optoelectronic alternative (99.460%). Noting disparate costs, the authors conclude: 'this result shows that it is possible to develop music-gesture learning applications based on low-cost technology, which can be used in home environments for self-learning practitioners'. Detailing the performance of an Hidden Markov Model (HMM)-based bow-articulation classification implementation, Sarasúa et al. (2017) noted enhanced early recognition rates when including Myo electromyography data; however, higher overall gestural recognition rates were achieved using IMU data alone.

Assessing optoelectronic motion capture, D'Amato et al. (2020) approached participant skill level identification using random forest classification algorithms, reporting classification accuracies of 73.34% and 80.16% in binary leave-one-person-out and leave-one-exercise-out implementations. The authors report a relatively high frequency of misclassification for two of four expert violinists, attributing this to comparatively expressive gestural execution. Similarly, the authors report a very high frequency of misclassification for one of three beginner violinists (Accuracy: 10.36%). The authors suggest this instance was due to a minimally expressive gestural execution, attributing this to further concentration required of the participant given their limited ability. The authors acknowledge the effective functionality of the developed implementation: a classifier of 'fragility, uncertainty and hesitation' – traits inferred by the algorithm as indicative of lesser skilled participants, although evidently confounded by idiosyncratic participant motor strategies. The authors later suggest the prospective utility of deep learning models towards such classification analyses, deeming these 'more sophisticated tools'.

TELMI[2] project authors Volpe et al. (2017) acknowledged barriers to adoption of their research products. D'Amato et al. (2021) later assessed low-cost alternatives to the multi-camera array used in TELMI project analyses. The authors achieved a classification accuracy of 87.85% through use of the Kinect – a low-cost computer-vision device produced by Microsoft Corp. (Redmond, WA, USA). While its success proved significantly lower than that of the benchmark optoelectronic system (96.98%), the authors reported that the highest classification accuracy was achieved through use of the Myo (98.15%). Following discontinuation of the Myo in 2018, an assessment of the Apple Watch (Apple Inc., Cupertino, CA, USA) as a comparable alternative for the purposes of violin performance analyses was detailed previously in Wilson et al. (2023).

In practice, the adoptability of any TEL implementation is bound by the interpretability of its output. While the ability to deduce macro characteristics, such as level of expertise, may facilitate observation or charting of user ability and progression, this offers limited constructive feedback by which future practice may be informed. Various authors have sought to address this, quantifying inferred quality indicators like bow skewness, dynamic stability and pitch stability through feature derivation.

Assessing complete beginners, Blanco et al. (2021), evaluated SkyNote, a TELMI-developed Graphical user interface (GUI), for the provision of real-time visual feedback (RTVF). Continually quantified, the three aforementioned performance metrics were displayed to participants during performance. The authors indicated mixed results.

During a final 'retention' condition, the authors noted occasionally significant improvements to performance metrics – albeit following 40 'baseline' and 'acquisition' trials. During intermediate 'acquisition' trials, RTVF was found to impede participants' sound-quality metrics. Between 'baseline' and 'retention' conditions, the experimental group demonstrated greater improvements. However, the intra-group improvement between initial 'acquisition' (sans RTVF) and 'retention' conditions appears far less pronounced. These findings suggest retained improvements were primarily a result of short-term repetition versus RTVF. The authors also acknowledge RTVF may pose a distraction during practice; 55% of participants polled agreed with the statement: 'Cannot play whilst watching feedback'. While Provenzale et al. (2024) also reported a reduction in bowing errors during assessment of their own RTVF system, 10 of 12 participants reported difficulty interpreting feedback while playing. While the real-time aspect of such implementations may seem inherently advantageous, these findings indicate that the feasibility of such feedback may be constrained as a result of additional user attention demands. For this reason, we consider the real-time availability of data to be a non-crucial functionality at this time.

2.3 AUDIO FEATURE EXTRACTION

Schedl et al. (2014) declare the aim of music information retrieval (MIR) techniques to be 'the extraction and

inference of meaningful features from music'. Towards this aim, low-level descriptors are used to quantify characteristic aspects of audio signals. Occasionally, such descriptors may prove visually interpretable—as in a spectrogram, wherein note durations, frequencies, and harmonics can be observed. Elsewhere, the primary utility of such descriptors is their suitability towards integration within automated classification systems.

Waveform representations of audio data depict signal amplitude over time. Frequency–domain representations differ, depicting constituent frequency magnitudes comprising the input signal. Such a representation can be calculated via discrete Fourier transform. One-dimensional, this representation fails to depict changes to the harmonic content of an audio signal over time. Schedl et al. (2014) suggest a short-time Fourier transform (STFT) 'spectrogram' for this purpose. The spectrogram includes iterative discrete Fourier transformations, denoting averaged harmonic content for short, successive 'frames' of an input signal.

Defined as 'the results of a cosine transform of the real logarithm of the STFT expressed on a mel-frequency scale' (Zheng et al., 2001), mel-frequency cepstral coefficients (MFCCs) have demonstrated utility within speech-recognition systems (Davis and Mermelstein, 1980). Providing a 'compact representation of the spectral envelope' (Wu et al., 2017), MFCCs have demonstrated similar utility in MIR applications. These include artist and genre identification (Li and Ogihara, 2004; Mandel and Ellis, 2005) and violin bow-stroke classification (Alar et al., 2021).

While useful for representing timbre, McFee et al. (2015) consider these limited in their depiction of pitch, offering 'poor resolution of pitches and pitch-classes'. Given an ability to 'encode harmony while suppressing variations in octave height, loudness or timbre', the authors instead suggest use of Chroma representations towards the depiction of pitch.

2.4 DNNs

Comprising numerical inputs fed towards a single node, Alpaydin (2020, p. 201) declares the perceptron: 'the basic processing element' of any DNN. Weights are ascribed to each numerical input; an output value is produced through summation of the product of each input and ascribed weight. The multi-layer perceptron (MLP) may be considered an expansion of the single-layer perceptron, containing any number of node layers linked by interconnecting weights feeding forward towards an output layer. During training upon labelled data, weights are refined, towards classification of input data to a corresponding output. The utility of such networks hinges on their ability to generalise, towards accurate classification of unseen data.

A development of the MLP, the recurrent neural network (RNN) incorporates weights fed backwards, from the output of intermediate nodes to the input of preceding nodes. Such connections enhance the ability of the network to interpret sequential contexts of input data, rendering these well suited to time-series classification tasks (Russell and Norvig, 2020, p. 824). Developed further, the bidirectional RNN processes input data in both chronological and reverse-chronological order simultaneously, facilitating classification based on both 'past' and 'future' information. Insignificant temporal associations are disregarded through gating of recurrent connections; devised implementations vary in complexity. Summarising two such implementations (Chung et al., 2014, p. 5) distinguish between gated recurrent units (GRUs) and long-short term memory (LSTM) units – attributing enhanced capability of the latter, to control the quantity of stored memory-content, to an incorporated third gate.

Ismail Fawaz et al. (2019) liken one-dimensional convolutional neural networks (CNNs) to the application of a sliding filter over a time-series. Discussing applications of these towards time-series classification tasks — including detection of cardiac arrhythmias and abnormal structural vibrations—Kiranyaz et al. (2021) conclude: '...even a low-power mobile device [...] will suffice to make real-time monitoring and analysis possible'. Detailing the incorporation of one-dimensional CNNs within multi-input deep CNNs (MI-DCNNs), Chen et al. (2022) characterise these as multiple parallel CNNs fed, via concatenation, into an MLP. During assessment of such networks towards classification of bioinformatic data, the authors reported higher classification accuracies compared to conventional machine-learning architectures.

Neural architectures detailed within this section were selected for discussion given their use during forthcoming primary analyses.

3 MULTIMODAL DATA CAPTURE AND SEGMENTATION

This section recounts a compilation of the primary dataset used. While detailed previously in Wilson et al. (2023), an understanding of this process is considered essential towards contextualising subsequently detailed analyses. A comparison of existing datasets is depicted in Supplementary Table 1, including summaries of their use in prior analyses.

3.1 RECORDING

Participants were sourced from the student body of the Royal Birmingham Conservatoire, each having received substantial, formal conservatoire training in classical violin performance. Research participation was not incentivised. All recordings were conducted in a controlled Conservatoire live room, characterised by a low noise floor and short reverberation time. This sought to minimise

room-acoustic colouration of the recorded audio, while remaining a setting familiar to research participants.

Nine participants (three male, six female) were asked to recite two-octave G and D Major scales, repeating every note twice such that an up- and a down-bow value was captured for each. Each violinist recited these using their own violin. These exercises were selected given a reasonable expectation that they would prove familiar to all research participants. Further, the consistency of defined tonal and temporal 'targets' was considered utile towards subsequent observation of performer tendencies. Participants performed each scale in two bow articulations: *Legato* and *Spiccato*. Three takes of each exercise were captured. Participants were communicated a target tempo of 110 BPM, with a duration of one bow-stroke per beat. All participants were provided a reference metronome initially, and upon request; however, participants were not permitted to use a metronome during active recording.

Participants wore an Apple Watch Series 8 (model no. 2770), running VioLogger, an ad-hoc recording interface, based upon Logger7[3] by GitHub user Shakshi3104. VioLogger synchronously logs both Audio (SR = 44.1kHz) and IMU (SR = 100 Hz) data, captured via a device's microphone and accelerometer. Audio files are timestamped upon the commencement and cessation of recording; IMU samples are timestamped individually.

3.2 DATA SEGMENTATION

Low-pass filtered root mean squared (RMS) envelopes were used to gate audio recordings, towards the removal of excess audio, pre- and post-activity. A threshold of 0.6× the envelope mean was calculated; indices at which the filtered envelopes first exceeded the calculated threshold were used to gate the start of the recording. Similarly, indices at which the filtered RMS envelopes finally subceded the calculated threshold were used to gate the end of the recording. Sequential application of the prior method at cutoff frequencies of 0.5 and 2 Hz were found to effectively remove soft-onset and transient noise, respectively.

IMU data concurrent with the trimmed audio data were identified via timestamping; excess data beyond the bounds of the trimmed audio duration were discarded. Resultant multimodal recordings were segmented into series of component bow-strokes through use of the Madmom bi-directional RNN onset detector.

A total 3455 segments of multimodal data were isolated; this is lower than the expected 4176 notes (58 notes × 3 takes × 6 participants × 2 scales × 2 articulations), constituting 83% of the expected total. This figure approximates the accuracy (F1: 0.778) observed by Tomczak and Hockman (2023) during application of the Madmom detector towards their QTDS string recordings dataset. Reported findings indicated greater precision (0.926) versus recall (0.679), suggesting a greater prevalence of false negatives relative to false positives. Previously, this diminished sensitivity has been attributed to string instruments' characteristically 'soft onsets' (Bock and Widmer, 2013). During implementation, such inaccuracy produces segments comprising multiple notes. To mitigate, segments of a duration considered infeasible (>0.8 s / <75 BPM) were excluded from classification. The number of segments isolated corresponds to a mean of 575.8 per participant[4]. Segments used in classification totalled 27 min and 10 s. Each comprising audio and IMU data, these averaged 0.47 s in duration (standard deviation: 0.09 s). While this approximates an expected average duration of 0.54 s, the shorter mean-duration indicates participants exceeded the target tempo by an average of 17.7 BPM.

4 AUDIO AND IMU DATA PROCESSING

The extent of processing required prior to the forthcoming analyses varied by datatype. Where implemented, data processing sought to ensure the suitability of recorded data towards classification and statistical analyses. Although recorded simultaneously via our multimodal smartwatch interface, IMU and Audio data were processed separately.

4.1 AUDIO DATA PROCESSING

Features were derived from isolated audio segments, towards a compact representation. These sought to minimise computational expense of classification while maintaining recorded audible characteristics. Features selected sought to represent recorded audio segments comprehensively, towards time-series representation of both pitch and timbre.

Following demonstrated utility during prior classification tasks, MFCCs were selected as a preferred feature for the representation of timbre, given their compact resolution and precedented use in prior participant recognition analyses (Ali-MacLachlan et al., 2020). MFCC matrices of sizes $13 \times M$ were computed (13 is the number of cepstral coefficients computed per frame; M is the number of frames). During computation, a Hanning window of 2048 samples was used, alongside a hop length of 256 samples. Similarly sized matrices of Delta- and Delta-Delta-MFCCs were computed. Concatenated along their first axis, these matrices formed a resultant matrix of $39 \times M$).

Towards representation of pitch, Chroma coefficients were selected as a preferred feature, computed via the Librosa fixed-window STFT-based implementation discussed in McFee et al. (2015, p. 20). To maintain concurrency with prior calculated MFCCs, these were framed similarly. Derived Chroma coefficients were subsequently

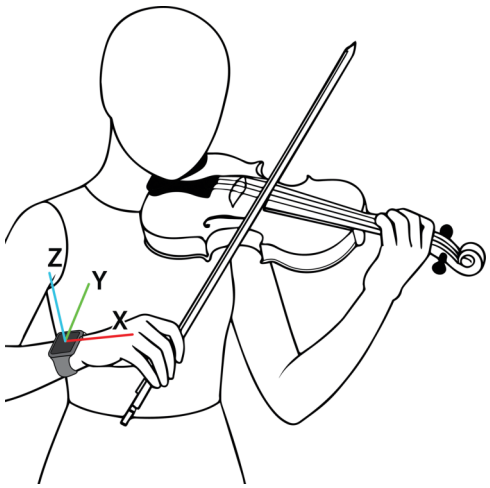

Figure 2 Apple Watch rotational axes as relative to the device.

concatenated with the initial MFCC matrix, such that each column comprised the four features per frame.

4.2 IMU DATA PROCESSING

Given a comparatively low sample rate, recorded IMU data necessitated no such feature derivation. Recorded via the Core Motion API[5], 'Device Motion' data were favoured versus raw IMU data. Normalized relative to device geometry, as depicted in Figure 2, this enhanced interpretability during plotting and indicated greater accuracy during preliminary testing for the identification of bow direction, due largely to a visible reduction in noise. Axes depicted in Figure 2 correspond to gyroscopic axes depicted in Figure 1.

5 NEURAL NETWORK–BASED PARTICIPANT RECOGNITION

Given an intention to gauge performer idiosyncrasy, participant recognition was implemented via neural network classification, towards an assessment of the extent to which recorded data proved identifiable.

5.1 NEURAL NETWORK IMPLEMENTATION

To facilitate an assessment of multimodality as a tool to enhance accuracy in violin note classification, a variety of disparate DNN architectures were used in classification.

Conventional sequential DNNs were used towards the classification of unimodal input datatypes, comprising a single input layer (Figure 3A), two hidden-layers (Figure 3B) – of types denoted by the Network Architecture column in Table 2 – followed by a densely connected layer (Figure 3E) and an output layer (Figure 3F).

Comprising individual subnetworks similar in form to the aforementioned single-input networks, MI-DNNs were used, enabling classification of both audio and IMU datatypes within a single, unified architecture. Final dense layers of these unimodal subnetworks were

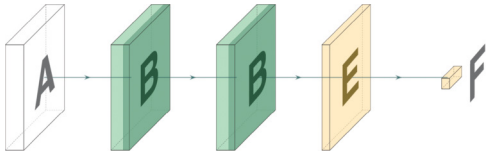

Figure 3 Conventional deep neural network used in unimodal classification implementations, comprising an input layer (A), two variable layers (B), a dense layer (E) and an output layer (F).

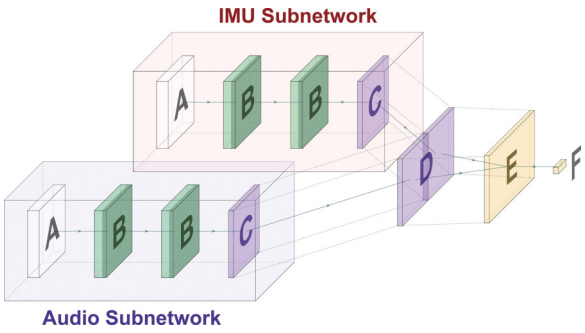

Figure 4 Multi-input deep neural network used in multimodal classification implementations, comprising two input layers (A), four variable layers (B), two flattened layers (C), a concatenation layer (D), a dense layer (E) and an output layer (F).

Layer	Parameter	MLP	GRU	LSTM	CNN
A	Units	128	–	–	–
	Activation	ReLU	–	–	–
B	Units/Filters*	128 128	128 88	128 88	50* 50*
	Kernel Size	–	–	–	5
	Activation	ReLU	ReLU	ReLU	ReLU
C	Units	88	–	–	–
	Activation	ReLU	–	–	–
D	N/A	–	–	–	–
E	Units	64	–	–	–
F	Units	N Classes	–	–	–
	Activation	Softmax	–	–	–

Table 1 Participant Recognition Network Parameters.

flattened (Figure 4C), concatenated (Figure 4D) and fed through a further dense layer (Figure 4E) to an output layer (Figure 4F).

A stratified 80:10:10 train/test/validation split was used alongside five-fold cross-validation, enhancing the reliability of model behaviour observations. An early stopping function halted training following failure to reduce validation loss over eight successive epochs. Once stopped, model weights were reset to those that had produced the lowest observed validation loss before testing. Network parameters used are depicted in Table 1.

5.2 RESULTS

Table 2 depicts test classification metrics – comprising classification accuracy, areas under the receiver operating characteristic curve, and F-score values – by architecture and input datatype. Proving the most accurate classifier, the single-input LSTM trained upon lone IMU data demonstrates that combining data modalities may not always reliably improve classification accuracies.

Figure 5 depicts train/validation accuracies calculated per-epoch, per-fold. These suggest MI-DNN architectures exhibited a greater degree of overfitting when compared to the Single input (SI)-DNN IMU classifier; this is likely due to an increased number of model parameters. SI-DNN Audio networks demonstrated the greatest degree of overfitting, however, as is demonstrated by the divergence of train/validation accuracies. Notably, SI-DNN Audio training accuracies also demonstrate a degree of underfitting relative to thier IMU-trained counterparts. This suggests the idiosyncrasy of participants' gestural execution may exceed that of their audible output. However, this result is also a product of feature selection, reflecting the utility of selected features; it is likely that alternative feature selection would influence classification accuracies – thus, the prior interpretation cannot be assumed to be correct.

Networks trained upon IMU data – whether combined with derived audio features or otherwise – exhibited higher classification accuracies. Assessed via a *t*-test, this proved to be statistically significant at $P < 0.01$. Neither grouping indicated a significant departure from normality when assessed using a Shapiro–Wilk test (IMU: $P = 0.415$, Audio: $P = 0.998$). While achieving a higher mean accuracy over tested network architectures (94.67%) than their MI-DNN counterparts (93.88%), single-input IMU classifiers did not demonstrate a statistically significant improvement in classification accuracy ($P = 0.426$). Shapiro–Wilk testing again indicated no significant departure from normality amongst these groupings (SI-DNN: $P = 0.134$, MI-DNN: $P = 0.290$). Despite the implications of this statistical testing, while normally distributed, notable variation in classification accuracy was observed between single-input IMU networks (91.83–96.43); single-input IMU LSTM and CNN1D classifiers achieved the highest classification accuracies observed (96.43 and 96.00, respectively), with the former demonstrating a +0.84% increase in classification accuracy versus the most accurate MI-DNN (GRU-GRU). This finding demonstrates that the inclusion of an additional modality cannot be assumed to enhance classification accuracy in all instances, despite further available data upon which to train.

Input Data Type	Network Architecture		Participant Recognition		
			Acc (%)	AUC	F-Score
Audio	MLP		76.38	0.948	0.765
”	LSTM		82.43	0.967	0.828
”	CNN1D		80.75	0.964	0.812
”	GRU		79.39	0.957	0.796
IMU	MLP		94.41	0.991	0.947
”	**LSTM**		**96.43**	**0.996**	**0.965**
”	CNN1D		96.00	0.994	0.961
”	GRU		91.83	0.989	0.918
Audio†	MLP†	MLP	94.66	0.993	0.950
+	LSTM†	LSTM	93.85	0.988	0.938
IMU	CNN1D†	CNN1D	94.20	0.986	0.942
”	GRU†	GRU	95.59	0.995	0.956
”	MLP†	CNN1D	94.08	0.992	0.942
”	CNN1D†	MLP	93.68	0.992	0.937
”	LSTM†	CNN1D	93.79	0.987	0.936
”	CNN1D†	LSTM	91.72	0.979	0.911

Table 2 Participant recognition classification accuracy metrics, by network architecture and input data type.

†denotes corresponding subnetwork modality.

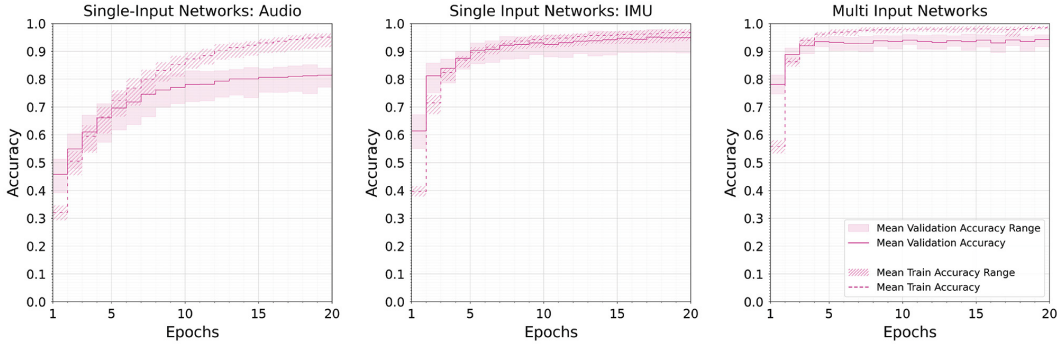

Figure 5 Participant train/validation accuracies per fold, by epoch, averaged across networks by datatype.

5.3 DISCUSSION

The ability of such architectures to distinguish participants suggests the recorded data to be inherently idiosyncratic – incorporating unique, identifying characteristics of participants' execution. Further, enhanced accuracy through the inclusion of gestural data suggests that such idiosyncrasies are not solely audible in nature. This implementation demonstrates that – while the combined use of multiple modalities cannot be assumed to be invariably beneficial – the use of additional modalities available may serve to enhance conventional MIR implementations and may prove feasible through the use of commercially accessible hardware. Within the context of MIR, this could be considered an application of the advice of Bugaj et al. (2019): to consider music through movement. Notably, the aforementioned definition attributed to Schedl et al. (2014) does not compel the use of audio as a medium for the study of MIR.

Each architecture discussed may be considered a black box; despite demonstrating that distinctions between participants prove identifiable, such networks' 'reasoning' is minimally interpretable. Thus, we consider conventional statistical analyses utile in observing and understanding such differences.

6 TENDENCY ANALYSES

Following prior demonstration of participant idiosyncrasy via DNN classification, statistical analyses seek to explore the nature of such idiosyncrasy. These quantify tonal and temporal performer tendencies, both individual and cross-participant.

6.1 TEMPORAL AND TONAL DISTRIBUTION ANALYSES

Towards the provision of interpretable practice feedback, the visualisation of participant tendencies was considered extensively. Intended to further interpretability of resultant data, calculated note durations were converted to their equivalent BPM. Note tunings were calculated in cents using the yin pitch-tracking algorithm[6]; deviation of a recorded note's detected fundamental frequency is calculated from the frequency of the nearest note established within equal temperament conventions (A4 = 440 Hz).

Figure 6 depicts temporal execution at the participant-level, via kernel density estimate plotting. This displays the distribution of note lengths for each participant in G Major *Legato* and *Spiccato* conditions. Upper, red-toned density curves depict the distribution of up-bow note durations. Conversely, lower, blue-toned density curves depict the distribution of down-bow note durations. While the shaded, left-most portion of the graph is impossible to play (given the target tempo, playing below this region would require notes of negative durations), this region has been retained, allowing plotting of the target tempo centrally along the X-axis, such that slow/fast = left/right. Each curve includes three shades, representing the distribution of note durations in each take. Darker, outermost curves depict initial takes, while lighter, innermost curves depict final takes. These are stacked, such that the innermost curves effectively form the X-axes of each successive curve. This seeks to facilitate observation of overall participant trends, while maintaining visibility of iterative changes in participant execution. Figure 7 is composed similarly, albeit with assumed target tuning (0 cents) plotted instead along the Y-axes. Tuning plots were transposed in this manner given conventional vertical depictions of pitch in staff notation.

Participants exhibited greater temporal variance in the *Spiccato* condition, indicating lesser temporal precision. During post-hoc consultation, participants attributed this to the inherently disjointed nature of such bow-strokes compared to the *Legato* condition, citing lesser continuity to provide timing references for sequential bow-strokes. Participants also consistently exhibited higher

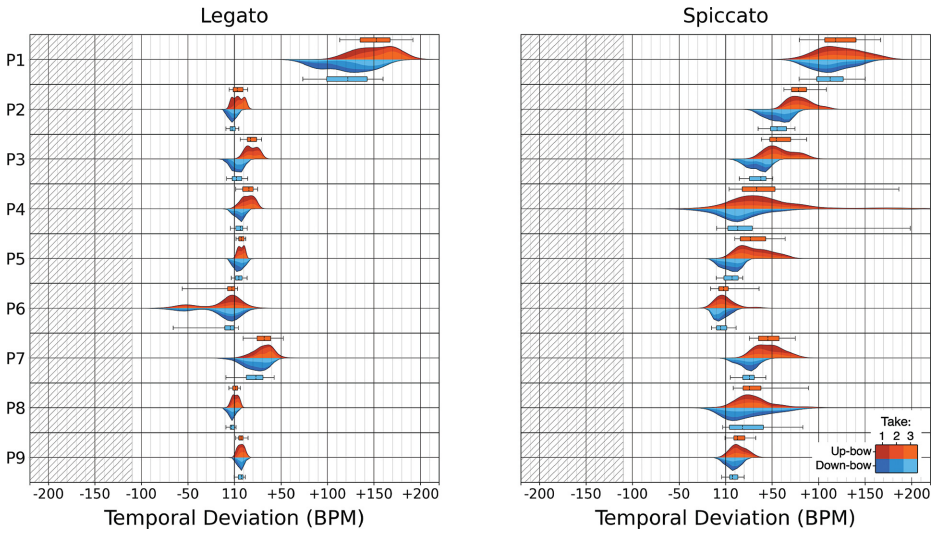

Figure 6 Distribution of participants' G Major scale individual note tempi, by articulation.

		Tonal Deviation (cents)			Temporal Deviation (BPM)		
	Bow Condition	**Down**	**Up**	**Both**	**Down**	**Up**	**Both**
Mean	Legato	0.48	4.40	2.43	8.97	15.90	12.42
	Spiccato	−6.16	−4.19	−5.17	33.64	53.03	43.35
	Both	−3.04	−0.19	−1.62	22.07	35.72	28.89
Std.	Legato	12.03	11.70	12.03	23.78	23.44	23.86
	Spiccato	15.03	14.95	15.20	52.50	57.08	55.68
	Both	14.10	14.19	14.22	43.35	48.35	46.42
Variance	Legato	144.8	136.6	144.5	565.3	549.3	569.1
	Spiccato	225.9	223.4	225.5	2756.2	3257.9	3100.5
	Both	198.8	201.3	202.1	1879.6	2337.6	2154.6
Median	Legato	1.90	5.30	3.70	6.02	12.22	8.80
	Spiccato	−4.30	−1.90	−3.10	18.60	38.29	27.89
	Both	−1.40	2.00	0.60	9.28	19.64	14.21
Kurtosis	Legato	1.613	2.064	1.650	37.37	29.61	31.87
	Spiccato	0.510	0.884	0.658	5.301	1.631	2.866
	Both	1.080	1.471	1.199	10.41	4.743	6.755
Skew	Legato	−2.16	−0.49	−0.347	5.316	4.527	4.767
	Spiccato	−0.50	−.722	−.608	2.305	1.431	1.777
	Both	−0.528	−.770	−.638	3.075	2.152	2.514

Table 3 Tonal and temporal deviation descriptive statistics, by bow articulation and direction.

average tempo in the *Spiccato* condition, likely stimulated by the rapid acceleration necessitated by this articulation. While interpretation of the aforementioned as indicators of 'quality' or level of expertise may seem intuitive, given our implementation of quantitative techniques in an inherently qualitative field, we do not seek to assess participants in this way.

Given a visibly anomalous temporal distribution, Participant 1 was excluded from subsequent statistical significance and descriptive testing. A total 5643 datapoints remained, across eight participants. Participant-specific anomaly detection was implemented via computation of z-scores. Datapoints with z-scores greater than 3 – indicating three standard deviations from the mean – were discarded; these amounted to 251. Given resultant sample sizes of between $N = 586$ and $N = 712$ per participant, Shapiro–Wilk tests were used to assess normality; results indicated both BPM and Tuning deviations are non-normally distributed in Articulation and Bow Direction conditions, with significance values in all cases of <0.001. Appropriately non-parametric, Wilcoxon signed-rank tests were used to assess the significance of observed differences in the grouped distributions. These returned *P*-values of <0.001, indicating significance.

Table 3 presents statistical descriptors averaged across the included eight participants – from these, a number of aforementioned trends can be quantified. Participants played the scale exercises faster than the established target tempo in both *Legato* and *Spiccato* conditions, averaging 138.89 BPM (+26.26%) across both up- and down-bows. Following discussion with participants during recording, it is expected that familiarity with such fundamental exercises led to this overall trend. Notably, the highest deviation in mean tempo was observed in the *Spiccato* condition, wherein participants averaged 153.35 BPM (+39.41%). Given sample sizes of $N = 2996$ and $N = 2647$, a Kolmogorov–Smirnov test was used to confirm normality, returning a *P*-value of <0.001. An independent-samples *t*-test was used to assess significance of the observed trend, again returning a *P*-value of <0.001. Likewise, participants were observed to perform up-bows at a higher tempo than down-bows, averaging these 13.65 BPM (+10.34%) faster. Again, Kolmogorov–Smirnov testing indicated normality ($P < 0.001$); a subsequent independent-samples *t*-test indicated statistical significance ($P < 0.001$).

This finding indicates that a performer's effective tempo is the product of averaging across unequal

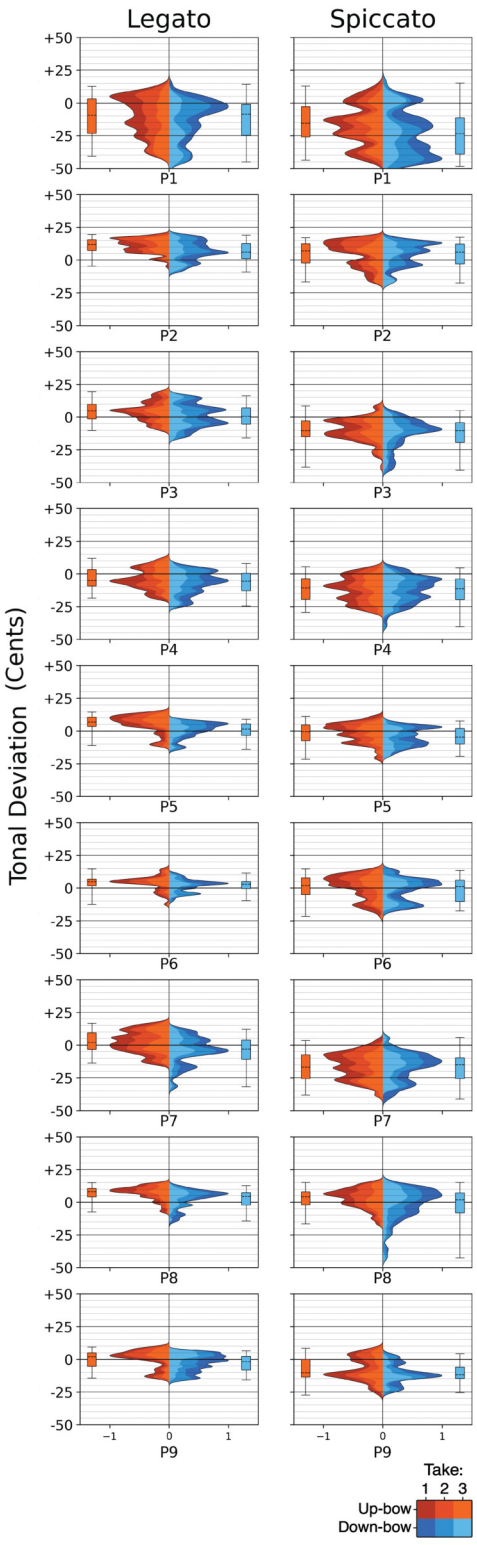

Figure 7 G Major scale tonal distributions per participant, by articulation.

up– and down–bow-stroke durations; respectively, these are observed as characterised by relatively short and long durations.

Through further consideration of the aforementioned means, a compounding effect of these two factors can be seen; the highest mean temporal deviation observed was in upwards *Spiccato* bow-strokes, at 53.03 BPM (+48.21%) faster than the target tempo. While participants exhibited similar temporal variance across both up-bows and down-bows, an increase of 2531.4 is observed in the *Spiccato* condition, corroborating the lesser temporal precision depicted previously. A Levene's test value of $P < 0.001$, indicates significance of this observed trend in variance. Kurtosis values indicate the greatest temporal precision in *Legato* down-bows. A moderate decrease is apparent through comparison to *Legato* up-bows; however, the most notable decrease in kurtosis is again observed as a result of the *Spiccato* condition. These trends appear to compound, in a similar pattern to mean temporal deviations. Similarly, the *Spiccato* condition appears to be the most prominent driver of tonal variance. Averaged variances differ little between up- and down-bows, and Levene's test indicates homogeneity of variance ($P = 0.744$). An increase of 81 is observed in the *Spiccato* condition, however. Corresponding standard deviations suggest tonal spread is increased by 3 cents, each side of the mean; this is substantiated by a Levene's test *P*-value of <0.001. We note the limitations of idealising equally-tempered pitches during tonal analyses; depending on musical context, such exact tunings may be neither preferred nor intended. Despite this, variation analyses remain utile in observing differences across conditions, as these are relative to performer mean tunings.

6.2 TEMPORAL AND TONAL CORRELATION ANALYSES

Bivariate correlational testing was used to assess the relationship between temporal deviation and tonal precision; BPM deviations were binned into equal percentiles of 5%, tonal variance was then calculated per bin. Resultant Kendall's Tau (τb) of 0.780 and $P < 0.001$ indicate a strong positive correlation between BPM and tonal variance. This correlation persists in partial correlation testing, when controlling for participant ID (τb = 0.864, $P < 0.001$). Separate correlational analyses indicate that this correlation was stronger in the *Spiccato* condition (τb = 0.940, $P < 0.001$) versus the *Legato* condition (τb = 0.682, $P < 0.001$). Subsequent linear regression indicates a steeper correlation gradient in the *Spiccato* condition (TonalVar = 1.963 (BPM) + 134.786) versus the *Legato* condition (TonalVar = 1.613 (BPM) + 150.004). These results demonstrate that, as tempo increases, tonal consistency is impaired regardless of note articulation; however, this decreased tonal consistency is more pronounced – and more reliably so – during the *Spiccato* condition.

6.3 NOTE-BY-NOTE ANALYSES

Intended to further the observability of take-specific participant trends, further plots were developed. These depict various aspects of participant execution note-by-note. Figure 8 depicts calculated note tunings over time,

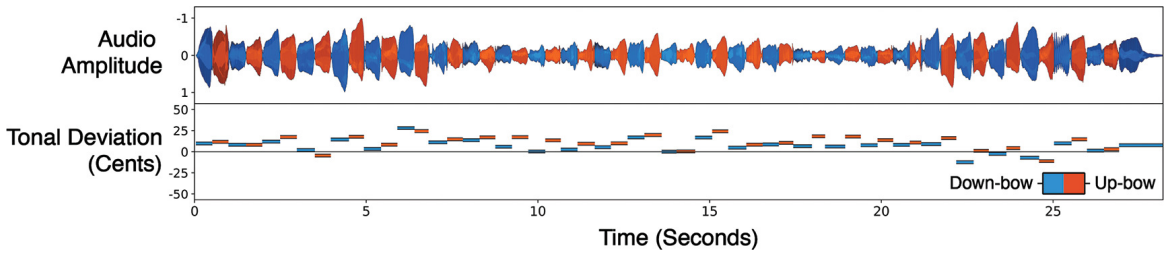

Figure 8 Note-by-note tunings, in cents, of a recorded D Major scale (*Legato*).

illustrating the constituent notes of which the prior distribution curves comprise.

Figure 9 depicts a piano-roll plot of a recorded D Major scale. Audio segments are aligned on a logarithmic Y axis, binned and labelled by equally tempered notes. Y-axis alignment is determined by the averaged fundamental frequency of each segmented note, calculated via the yin pitch-tracking algorithm. This presentation was favoured to enhance interpretability in prospective TEL use-cases, wherein familiarity with effectively similar plots such as the spectrogram cannot be assumed. Yellow-shaded regions denote expected note durations, demonstrating the effects of accumulated error due to even slight temporal inaccuracy.

6.4 NOTE FINGERING ANALYSES

In the G-major condition, Shapiro–Wilk tests again indicated BPM deviation and tonal variance were non-normally distributed per-note and per-finger. Wilcoxon signed-rank tests indicated significant differences between distributions of tonal variance and temporal deviation, at *P*-values of <0.001. Second finger notes demonstrated the highest tonal variance: 199.990; this contrasts with the lowest observation of 151.196, characterising open bow-strokes. This corresponds to an increased spread of +1.9 cents either side of the origin. Despite lesser precision, the second finger demonstrated the greatest tonal accuracy. Participants averaged a mean tuning of −0.7 cents, whereas open, first, and third finger notes averaged −2.44, −1.86 and −5.49 cents, respectively.

Similarly, significant differences were observed in the D Major condition. Participants averagely demonstrated a similar trend of slightly 'flattening' first, second and third finger notes (with respective mean tunings of −1.9, −0.56 and −0.24 cents). Fourth finger notes, in contrast, averaged notably sharp, at +2.17 cents. Tonal variance across fingers was reasonably consistent, ranging between 13.86 and 14.85 cents. Notes A3, F#4 and C5 demonstrated the lowest tonal precision, with respective standard deviations of 14.72, 14.53 and 14.93 cents. Notes A4 and C4 demonstrated the greatest tonal precision, with respective deviations of 10.45 and 10.88 cents. Lacking correspondence across conditions, these results do not suggest note-fingers to be a reliable predictor of habits in tonality.

6.5 SUMMARY

While it should not be assumed that observed participant trends ground the participant classification accuracies demonstrated previously, such analyses do indicate statistically significant idiosyncrasies shape tonal and temporal execution.

Immediately, one might conclude that performers may wish to expend greater attention towards the practice of *Spiccato* bow-strokes, addressing consistently lower degrees of temporal precision and accuracy. Likewise, one may infer that, through repeated practice,

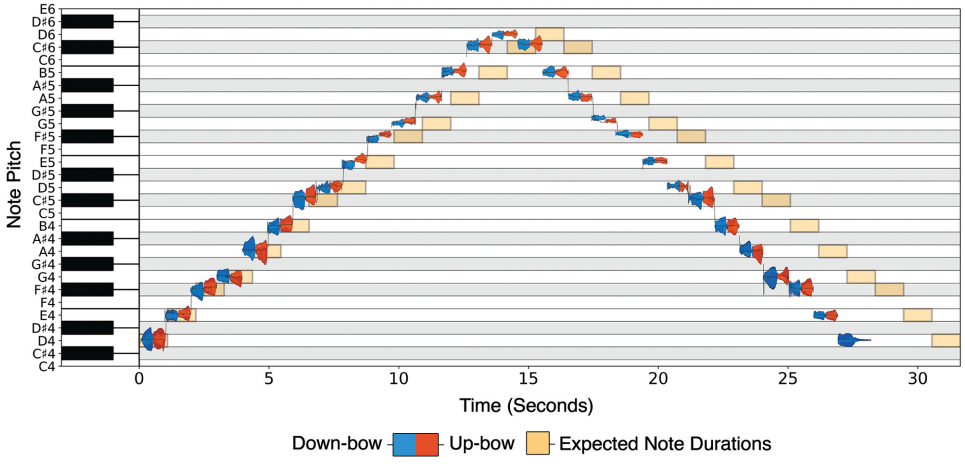

Figure 9 Piano roll graph depicting a recorded D Major scale (*Legato*).

performers may minimise their individual correlation between tempo and tonal variance, towards greater tonal consistency at higher tempi. In authoring our findings, however, we suggest that the role of biomechanical constraints – as discussed in prior literature – must be considered carefully before drawing such conclusions; almost certainly, task-specific ceilings exist at which further quantitative 'improvements' cannot be made. Identification of these would require further, longitudinal study; these are not identifiable within the scope of this work.

We must acknowledge that the application of such quantitative perspectives may be considered reductive in an inherently qualitative discipline. Notably, during discussion, one participant characterised the implied demands of such feedback as 'robotic'. It is therefore necessary to consider the applicability of an isochronous 'target' tempo. While Chander et al. (2022) demonstrate that rigid isochrony may not be invariably desirable – particularly during solo recital, as was stipulated – certainly there remain circumstances wherein an ability to maintain an established tempo is crucial; Ihas et al. (2023) identify violin pedagogues Samuel Applebaum, Louis Kievman and Simon Fischer as proponents of metronome practise, while Prynn (2018) declares the metronome: 'an essential tool for the preparation of ensemble work....' We consider our proposed implementation feasible as a supplement to metronome practise, facilitating assessment of continued adherence to notated tempi following departure from the metronome. We do not profess that our implementation serves to replace traditional student/teacher instruction. Instead, we believe the documented ability to quantify and visualise such performer tendencies can offer utility to both teachers and students – whether as a tool for the triage of student 'habits' or the provision of 'habit'-informed practise feedback. While rudimental, presently recorded exercises facilitated our analyses of two fundamental performance aspects – tonality and temporality. Analyses of music of further technical complexity may be facilitated via score-following. This would enable observation of tonal and temporal deviation from defined melodies, comprising notes of different lengths and ground-truth pitches.

Our usage of the Apple Watch eliminated the requirement of many control measures typically imposed in optoelectronic studies. These include regulation of clothing, lighting and positioning of the participant (Chander et al., 2022; McGuirk et al., 2022). Further, the devised recording interface requires minimal set-up, necessitating no consideration of camera placement, framing, reference pose or marker-placement.

Naturally, this solution does not fully parallel optoelectronic systems – movement captured is limited to the wrist. Whole-body gestural consideration is endorsed in pedagogy (Bugaj et al., 2019), and the significance of whole-body movement has been demonstrated by Chander et al. (2022). The wrist, however, constitutes a fulcrum of the bow-stroke: the essential, sound-producing motion (Palac, 1992). Thus, while an analysis of 'non-technical' performer movements may necessitate whole-body motion capture, we deem this non-essential for the provision of such insights, focussed currently towards the bow-stroke.

Further exploratory work was undertaken to assess the system's utility towards analysis of a trial-recorded melody. Temporal and tonal deviations presented challenges identifying which segments corresponded to scored notes. To address this, we calculated recorded-note 'centroids' (midpoints of frequency and duration, normalised), pairing these with scored-note 'centroids', based on the minimum total Euclidean distance with which all could be paired. Issues arose during instances wherein false negatives produced an unequal number of centroids to be paired. Thresholding was therefore introduced, enabling disposal of erroneous centroids. Determination of a suitable threshold is complicated by variation in pitch and tempo. Figure 10 depicts a prototype implementation, demonstrating circumstantial utility; however, we do not consider this approach suitably reliable for further analyses at this time. Given the difficulties encountered, we anticipate that score-based onset detection would prove more feasible for analyses of non-isochronous melodies.

We also recognise that intentional deviation from 'metronomical exactitude' may be preferred during such expressive musical contexts (Flesch, 1939a, p. 50; Galamian, 1964, p. 7). The dependence of such intentional temporal deviation upon musical context has been demonstrated previously via quantitative means (Chander et al., 2022; Huberth et al., 2020). Further, we note limitations towards the feasibility of attaining 'perfect intonation'; again, this concept is discussed within pedagogical literature (Flesch, 1939b, p. 20). Our implementation provides a means by which the extent of such deviation – whether preferred, intended or otherwise – can be observed.

7 CONCLUSION

By quantifying technically simplistic aspects of a user's practice, meaningful tendencies – or 'habits' – can be observed, providing insights which may prove constructive by informing practice. Through combination of the documented analytical approach with an accurate, gesture-based articulation classifier – as detailed in Wilson et al. (2023) – we intend that such insights may be automated for excerpts containing multiple bow articulations.

This research sought to minimise barriers to both research engagement and conduction, through use of a

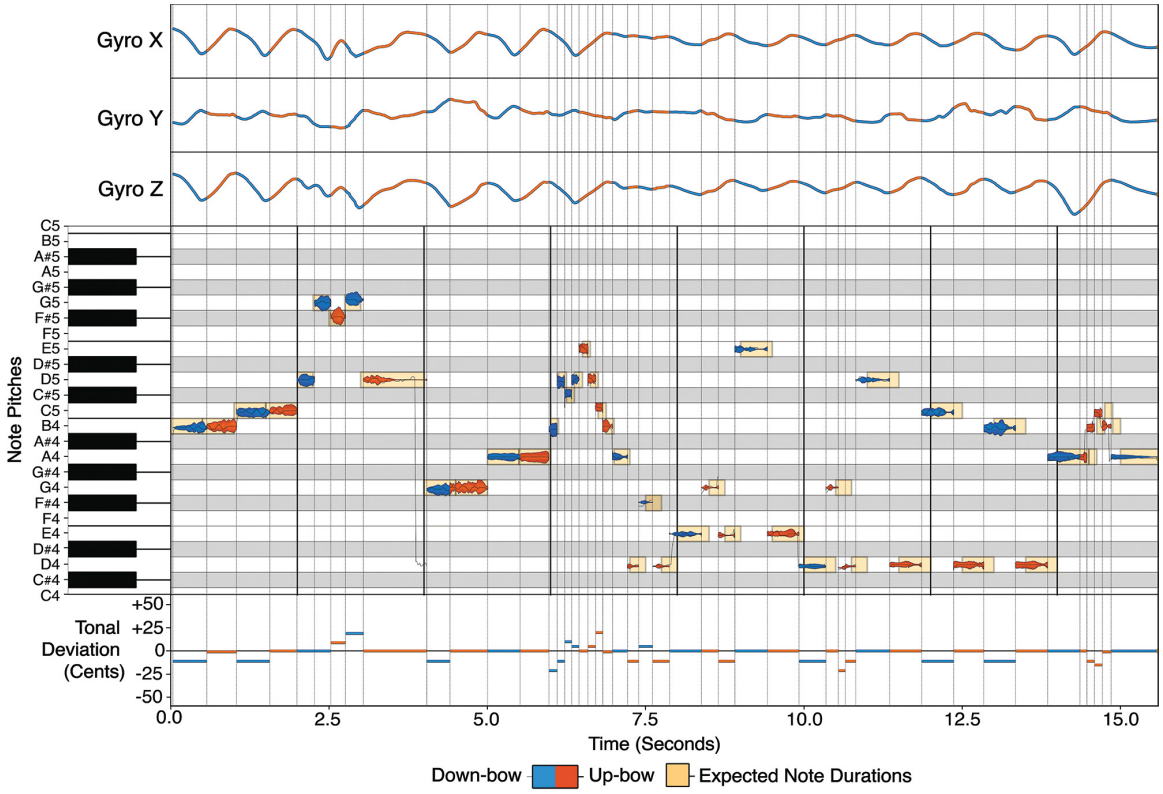

Figure 10 Piano roll graph depicting a multimodal recording of an excerpt from Gossec's *Gavotte*, bars 27–34. Tempo: *Allegretto* (120 BPM).

device more readily available than those used in prior violin TEL applications; however, the significance of remaining barriers should not be understated. Ownership of an Apple Watch cannot be assumed, limiting adoptability of our approach in real-world practice scenarios. Given observed, audible differences between up- and down-bows, an audio-only classifier facilitating identification of these may prove feasible. Use of available gestural modalities may enable development of this through expedited, albeit assumed labelling. Provided sufficient accuracy, this would provide similarly detailed practice feedback, while negating entirely the requirement of specific hardware.

ACKNOWLEDGEMENTS

Funding was received towards the completion of this research by Birmingham City University. We would like to acknowledge the valued contribution of research participants.

ETHICS AND CONSENT

This work was conducted after approval via Birmingham City University's research ethics procedures. Informed consent was granted individually, by all participants prior to participation.

AUTHOR'S CONTRIBUTIONS

Development of the recording interface, participant recording, data visualisation and subsequent analyses were conducted by W. Wilson. N. Granieri. provided significant contributions towards development of the recording interface and subsequent analyses. S. Smith provided valued guidance during statistical analyses. C. Harvey and I. Ali-MacLachlan contributed extensively towards developed methodologies.

COMPETING INTERESTS

The authors have no competing interests to declare.

DATA AVAILABILITY STATEMENT

Our dataset, used in the detailed analyses, is available at: https://github.com/willfwilson/smartwatch-based-audio-gestural-insights-in-violin-bow-stroke-analyses.

NOTES

1. https://madmom.readthedocs.io/en/latest/, last accessed: 03/08/25.
2. https://telmi.upf.edu, last accessed: 03/08/25.
3. https://github.com/Shakshi3104/Logger7, last accessed: 03/08/25.
4. At the time of the detailed classification experiment, six of nine total participants had been recorded.

5. https://developer.apple.com/documentation/coremotion/cmmotionmanager, last accessed: 03/08/25.
6. https://pypi.org/project/torch-yin/, last accessed: 03/08/25.

ADDITIONAL FILE

The additional file for this article can be found using the links below:

- **Supplementary File 1.** Comparison of Existing Datasets. DOI: https://doi.org/10.5334/tismir.216.s1.

AUTHOR AFFILIATIONS

William Wilson https://orcid.org/0009-0005-5332-2390
Acoustics and Audio Analysis research lab, College of Computing, Birmingham City University, Birmingham, B4 7RQ, UK

Niccolò Granieri https://orcid.org/0000-0002-0477-798X
Multimedia Interaction Designer, 4DODO S.r.l, San Giorgio dl Nogaro, 33058 Udlne, Italy

Samuel Smith https://orcid.org/0000-0001-9276-0354
Acoustics and Audio Analysis research lab, College of Computing, Birmingham City University, Birmingham, B4 7RQ, UK

Carlo Harvey https://orcid.org/0000-0002-4809-1592
School of Digital Arts, Manchester Metropolitan University, Manchester, M15 6ED, UK

Islah Ali-MacLachlan
https://orcid.org/0000-0002-9380-3122
Acoustics and Audio Analysis research lab, College of Engineering, Birmingham City University, Birmingham, B4 7RQ, UK

REFERENCES

Alar, H. S., Mamaril, R. O., Villegas, L. P., and Cabarrubias, J. R. D. (2021). Audio classification of violin bowing techniques: An aid for beginners. *Machine Learning with Applications*, 4, 100028.

Ali-MacLachlan, I., Hunt, E., and Jamieson, A. (2020). Player recognition for traditional Irish flute recordings using K-nearest neighbour classification. In *Proceedings of the 2nd International Conference on Timbre* (pp. 1–4). Thessaloniki, Greece.

Alpaydin, E. (2020). *Introduction to Machine Learning* (4th ed.). MIT Press.

Blanco, A. D., Tassani, S., and Ramirez, R. (2021). Real-time sound and motion feedback for violin bow technique learning: A controlled, randomized trial. *Frontiers in Psychology*, *12*, 648479.

Böck, S., and Widmer, G. (2013). Maximum filter vibrato suppression for onset detection. In *Proceedings of the 16th International Conference on Digital Audio Effects*, Maynooth, Ireland.

Bugaj, K. A., Mick, J., and Darrow, A.-A. (2019). The relationship between high-level violin performers' movement and evaluators' perception of musicality. *String Research Journal*, *9*(1), 23–33.

Chander, A., Huberth, M., Davis, S., Silverstein, S., and Fujioka, T. (2022). Violinists employ more expressive gesture and timing around global musical resolutions. *Music Perception*, *39*(3), 268–288.

Chen, P., Zou, B., Belkacem, A. N., Lyu, X., Zhao, X., Yi, W., Huang, Z., Liang, J., and Chen, C. (2022). An improved multi-input deep convolutional neural network for automatic emotion recognition. *Frontiers in Neuroscience*, *16*, 965871.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning*, Montreal, Canada.

D'Amato, V., Volta, E., Oneto, L., Volpe, G., Camurri, A., and Anguita, D. (2020). Understanding violin players' skill level based on motion capture: A data-driven perspective. *Cognitive Computation*, *12*(6), 1356–1369.

D'Amato, V., Volta, E., Oneto, L., Volpe, G., Camurri, A., and Anguita, D. (2021). Accuracy and intrusiveness in data-driven violin players skill levels prediction: MOCAP against MYO against KINECT. In *Advances in Computational Intelligence: 16th International Work-Conference on Artificial Neural Networks, IWANN 2021, Virtual Event, June 16–18, 2021, Proceedings, Part II*, volume 12862 of *LectureNotes in Computer Science* (pp. 367–379). Springer International Publishing.

Dalmazzo, D., Tassani, S., and Ramirez, R. (2018). A machine learning approach to violin bow technique classification: A comparison between IMU and MOCAP systems. In *Proceedings of the 5th International Workshop on Sensor-based Activity Recognition and Interaction* (pp. 1–8). ACM.

Davis, S., and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357–366.

Dippel, D. (1987). String bowings for the choral conductor. *The Choral Journal*, *28*(5), 11–13.

Flesch, C. (1939a). *The Art of Violin Playing: Artistic Realization and Instruction* (2nd ed., Vol. 2). Carl Fischer.

Flesch, C. (1939b). *The Art of Violin Playing: Technique in General Applied Technique* (2nd ed., Vol. 1). Carl Fischer.

Galamian, I. (1964). *Principles of Violin Playing and Teaching* (1st ed.). Faber and Faber.

Grimes, D. R. (2014). String theory: The physics of string-bending and other electric guitar techniques. *PLoS ONE*, *9*(7), e102088.

Guettler, K. (1997). Bow notes. In *Proceedings of the Institute of Acoustics* (Vol. 19). The Norwegian State Academy of Music.

Guettler, K. (2010). The bowed string. In **T. D. Rossing** (Ed.), *The Science of String Instruments*. Springer New York.

Helmholtz, H. (1895). *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Longmans, Green.

Huberth, M., Davis, S., and Fujioka, T. (2020). Expressing melodic grouping discontinuities: Evidence from violinists' rubato and motion. *Musicae Scientiae*, *24*(4), 494–514.

Ihas, D., Wilson, M., and McCormick, G. (2023). *Teaching Violin, Viola, Cello, and Double Bass: Historical and Modern Pedagogical Practices* (1st ed.). Routledge.

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., and Muller, P.-A. (2019). Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery*, *33*(4), 917–963.

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., and Inman, D. J. (2021). 1D convolutional neural networks and applications: A survey. *Mechanical Systems and Signal Processing*, *151*, 107398.

Li, T., and Ogihara, M. (2004). Music artist style identification by semi-supervised learning from both lyrics and content. In *Proceedings of the 12th ACM International Conference on Multimedia* (pp. 364–367).

Mandel, M. I., and Ellis, D. P. W. (2005). Song-level features and support vector machines for music classification. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, (p. 6). London, UK.

McFee, B., Raffel, C., Liang, D., Ellis, D., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference* (pp. 18–24). Austin, Texas.

McGuirk, T. E., Perry, E. S., Sihanath, W. B., Riazati, S., and Patten, C. (2022). Feasibility of markerless motion capture for three-dimensional gait assessment in community settings. *Frontiers in Human Neuroscience*, *16*, 867485.

Palac, J. (1992). Violin bowing technique: An analysis of contemporary pedagogical literature according to principles of human movement. *Medical Problems of Performing Artists*, *7*(1), 30–34.

Perez Carrillo, A. (2013). Characterization of bowing strokes in violin playing in terms of controls and sound: Differences between bouncing and on-string bow strokes. In *Proceedings of Meetings on Acoustics* (Vol. 19, pp. 198–212). Acoustical Society of America.

Provenzale, C., Di Tommaso, F., Di Stefano, N., Formica, D., and Taffoni, F. (2024). Real-time visual feedback based on MIMUs technology reduces bowing errors in beginner violin students. *Sensors*, *24*(12), 3961.

Prynn, G. (2018). Performing new music: Rhythms and meters in post-1950 repertoire. *American Music Teacher*, *67*(5), 16–20.

Rasamimanana, N., Bernardin, D., Wanderley, M., and Bevilacqua, F. (2007). String bowing gestures at varying bow stroke frequencies: A case study. In **M. Sales Dias, S. Gibet, M. M. Wanderley and R. Bastos** (Eds.), *Gesture-Based Human-Computer Interaction and Simulation* (pp. 216–226). Springer. https://doi.org/10.1007/978-3-540-92865-2_24.

Russell, S., and Norvig, P. (2020). *Artificial Intelligence: A Modern Approach*, (4th ed.). Pearson Education.

Sarasúa, A., Caramiaux, B., Tanaka, A., and Ortiz, M. (2017). Datasets for the analysis of expressive musical gestures. In *Proceedings of the 4th International Conference on Movement Computing* (pp. 1–4). ACM.

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

Shan, G., Visentin, P., Wooldridge, L., Wang, C., and Connolly, D. (2007). A frequency-based characterization of spiccato bowing in violin performance. *Perceptual and Motor Skills*, *105*(3), 1027–1051.

Thorn, D. S. D. (2018). Alto.Glove: New techniques for augmented violin. In *Proceedings of NIME '18* (pp. 334–339). University of Virginia.

Tomczak, M., and Hockman, J. (2023). Onset detection for string instruments using bidirectional temporal and convolutional recurrent networks. In *Proceedings of the 18th International Audio Mostly Conference* (pp. 136–142). ACM.

Volpe, G., Kolykhalova, K., Volta, E., Ghisio, S., Waddell, G., Alborno, P., Piana, S., Canepa, C., and Ramirez-Melendez, R. (2017). A multimodal corpus for technology-enhanced learning of violin playing. In *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter* (pp. 1–5). ACM.

Wilson, W., Granieri, N., and Ali-MacLachlan, I. (2023). Ubiquitous multimodality as a tool in violin performance classification. In *2023 4th International Symposium on the Internet of Sounds* (pp. 1–8). IEEE.

Wu, Y., Wang, Q., and Liu, R. (2017). Music instrument classification using nontonal MFCC. In *Proceedings of the 2017 5th International Conference on Frontiers of Manufacturing Science and Measuring Technology (FMSMT 2017)*, Atlantis Press.

Zheng, F., Zhang, G., and Song, Z. (2001). Comparison of different implementations of MFCC. *Journal of Computer Science and Technology*, *16*(6), 582–589.

TO CITE THIS ARTICLE:
Wilson, W., Granieri, N., Smith, S., Harvey, C., & Ali-MacLachlan, I. (2025). Smartwatch-Based Audio–Gestural Insights in Violin Bow Stroke Analyses. *Transactions of the International Society for Music Information Retrieval*, *8*(1), 283–299. DOI: https://doi.org/10.5334/tismir.216

Submitted: 19 August 2024 **Accepted:** 1 August 2025 **Published:** 4 September 2025