Ubiquitous Multimodality as a Tool in Violin Performance Classification

1st William Wilson
SoMA Group
DMT Lab
Birmingham City University
Birmingham, United Kingdom
0009-0005-5332-2390

2nd Niccolò Granieri
Otorhinolaryngology and Audiology,
Institute for Maternal and Child Health
IRCCS "Burlo Garofolo"
Trieste, Italy
0000-0002-0477-798X

3rd Islah Ali-MacLachlan
SoMA Group
DMT Lab
Birmingham City University
Birmingham, United Kingdom
0000-0002-9380-3122

***Abstract*—Through integrated sensors, wearable devices such as fitness trackers and smart-watches provide convenient interfaces by which multimodal time-series data may be recorded. Fostering multimodality in data collection allows for the observation of recorded actions, exercises or performances with consideration towards multiple transpiring aspects. This paper details an exploration of machine-learning based classification upon a dataset of audio-gestural violin recordings, collated through the use of a purpose-built smartwatch application. This interface allowed for the recording of synchronous gestural and audio data, which proved well-suited towards classification by deep neural networks (DNNs). Recordings were segmented into individual bow strokes, these were classified through completion of three tasks: Participant Recognition, Articulation Recognition, and Scale Recognition. Higher participant classification accuracies were observed through the use of lone gestural data, while multi-input deep neural networks (MI-DNNs) achieved varying increases in accuracy during completion of the latter two tasks, through concatenation of separate audio and gestural subnetworks. Across tasks and across network architectures, test-classification accuracies ranged between 63.83% and 99.67%. Articulation Recognition accuracies were consistently high, averaging 99.37%.**

***Index Terms*—datasets, neural networks, gestural analysis, computational musicology, violin, IMU sensors**

I. Introduction

Deep Neural Networks (DNNs) have long proved an effective means of classification for time-series data. Considering aspects of musical performance which may be recorded as such, the availability of audible and gestural performance content emerges. Occurring concurrently, each may be quantified through the use of suitable sensor technologies. While audio data can be recorded through the use of a microphone, there is a broader range of technologies that may be used for the quantification of gesture. Prior findings of MLP based classification demonstrated higher participant classification accuracies following the inclusion of gestural data, indicating that participant (i.e. violinist) performance distinctions exceed audibility alone [1]. Prior studies have typically cited the potential utility of audio-gestural approaches towards the development of music-education [2]–[4] and score-transcription tools [5] as a motivation. Specialist apparatus such as multi-camera arrays [6], and niche consumer devices such as the now-discontinued Myo [7]–[9], have demonstrated capability in such implementations. The feasibility of such research products is impeded, however, by access barriers such as cost and user-required technical-expertise [10].

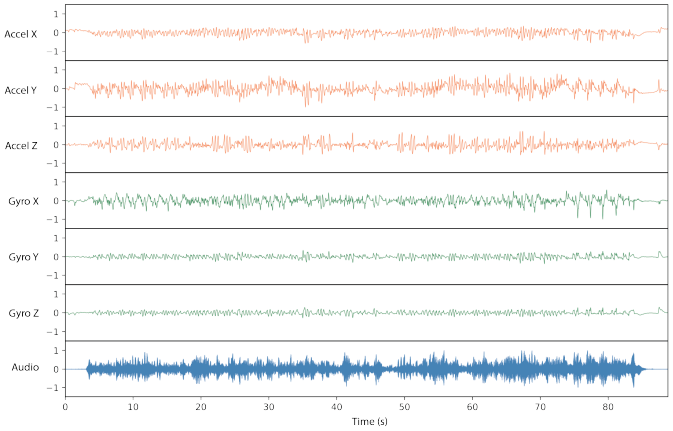

Fig. 1. Apple Watch IMU-Audio recording of Bach's Cello Suite No.1 in G Major, Prélude

Through the use of a comparatively ubiquitous device, such as the smartwatch, many of these barriers may be negated. Enhancements to the adoptability and longevity of research products have been observed through previous uses of such commonly accessible technologies. Investigating the adoption of novel Digital Musical Instruments (DMIs) presented at the *New Interfaces for Musical Expression* (NIME) conference, Morreale and McPherson [11] surveyed 70 prior NIME authors from 2010-2014. The authors found that 46.9% of the presented DMIs were in ongoing use, while just 40% had been played by more than three musicians. Of all DMIs included in the survey, the authors found that uptake was highest for iOS based applications; reported sales of these ranged from 1200 to 250,000. Excluding these, only three further DMIs had been sold to more than one buyer.

In literature, there exists no precedent for the use of such a mainstream wearable device - the smartwatch - towards the classification of violin bow-strokes. Based upon a primary corpus of musical scales, recorded through the use of an Apple Watch, we conduct three classification tasks towards an

assessment of multi-input approaches to the classification of multi-modal data. Participant classification accuracies indicate the ability of trained networks to identify performers through the learning of gestural and audible performer idiosyncrasy, while an articulation recognition condition indicates the utility of trained networks to disregard these, discriminating between two bowing techniques in a cross-participant implementation. A third task: Scale Recognition, seeks to investigate the ability of the networks to infer the belonging of a note to one of two scales, despite only a single note difference.

II. Background and Related Works

A. Gestural Sensors

In any three-dimensional space an object has both a location and an orientation; each can be described in three dimensions. The location of an object can be quantified through consideration of its translational position relative to a set of X, Y, Z axes. Likewise, the orientation of an object can be described by the object's rotation around each axis. These metrics are individually termed 'Degrees of Freedom' (DoF); a device quantifying both location and orientation in three dimensions would thus offer 6-DoF [12]. Inertial Measurement Unit (IMU) sensors record changes in each DoF over time, yielding three-axis acceleration and gyroscopic data.

Prior studies have assessed the efficacy of forearm-mounted IMU sensors during analyses of gestural execution in violin performance; Dalmazzo et al. [7], [13] and Sarasúa et al. [9] reported respective accuracy metrics of 0.98 (F1), 0.946 (F1), and 98.9% (Acc) when using the Myo Armband to capture IMU-series for the purposes of bow articulation classification. The latter authors reported increased early gestural recognition rates following the inclusion of Electromyography (EMG) data, although a decreased overall classification accuracy of 94.3% when compared to classification upon IMU data alone.

While assessing the feasibility of the Myo Armband as an alternative to optoelectric, marker-based motion capture (mocap) systems, Dalmazzo et al. reported higher classification accuracies through use of the former. Noting the relatively-contrasting costs of the two technologies, the authors propound 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" [8].

While assessing the viability of the Apple Watch as a tool for the purposes of hospital inpatient monitoring, Auepanwiriyakul et al. [14] employed a similar methodology, making use of the device for the purposes of activity classification via IMU logging. The authors concluded that "with relatively few drawbacks, consumer-grade smartwatches can be objectively used within a clinical- and research-grade setting", having compared the device to a "gold standard" optoelectronic OptiTrack system. While a similarly specialist optical mocap technology was implemented in the *Technology Enhanced Learning of Musical Instrument Performance* (TELMI) project towards the development of technology-enhanced learning interfaces, authors Volpe et al. [6] acknowledged a necessity for low-cost alternatives to their 12-camera array, for use by students and schools. Detailing Random Forest classification of participant skill level, D'Amato et al. [3] reported an accuracy of 87.85% through the use of such a low-cost computer-vision alternative: the Kinekt. In comparison, the authors reported accuracies of 96.98% through use of a Qualysis mocap system, and 98.15% through use of the Myo Armband.

D'Amato et al. [15] classified 7 participants with respective mean accuracies of 73.34% and 80.16%, through use of upper-body mocap and random forests. The authors reported an association between participant skill level and ease of classification.

B. Audio Feature Extraction Techniques

In summarising the aim of Music Information Retrieval (MIR) techniques, Schedl et al. [16] prescribe "the extraction and inference of meaningful features from music". Characteristic aspects of audio signals may be quantified through the calculation of established low-level descriptors. While some such descriptors prove visually interpretable, (insofar as relative frequencies, harmonics, and note durations may be inferred from a spectrogram - a depiction of the STFT), the utilities of many lie in their suitability towards integration within computational classification systems.

While time-domain representations of audio data denote the amplitude of an audio signal over time, frequency-domain representations may be used to depict the individual frequency magnitudes of which an audio signal is comprised. These may be calculated through computation of the signal's Discrete Fourier Transform (DFT). Such one-dimensional frequency-domain representations, fail to depict the temporal evolution of audio data. The authors suggest use of a Short-Time Fourier Transform (STFT) for this purpose, generated through iterative calculation of DFTs for short, successive 'Frames' of an audio signal [16].

The utility of Mel-Frequency Cepstral Coefficients (MFCCs) in speech recognition systems has been long demonstrated [17]. Through provision of a "compact representation of the spectral envelope" [18], MFCCs have proved similarly well suited towards applications in computational musicology, such as genre and artist-identification [19], [20], and violin bow stroke classification [21]. Zheng, Zhang and Song [22] define MFCCs as "the results of a cosine transform of the real logarithm of the STFT expressed on a mel-frequency scale"; a scale noted by Stevens [23] to approximate human auditory perception.

Despite their usefulness as a representation of timbrality, MFCCs are limited in their depiction of pitch, considered by McFee et al. [24] to offer "poor resolution of pitches and pitch-classes". Instead, for the depiction of these, the authors suggest use of Chroma representations, purporting these to "encode harmony while suppressing variations in octave height, loudness, or timbre".

C. Deep Neural Networks

Consisting of a single node and any number of numerical inputs, Alpaydin [25] identifies the perceptron as "the basic

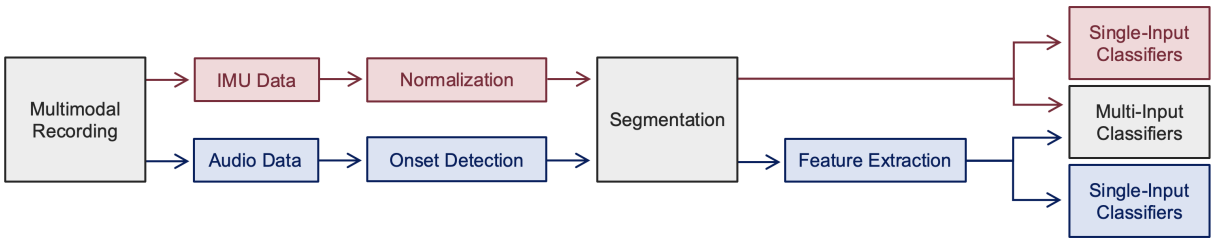

Fig. 2. Data Processing Flow Diagram

processing element" of a deep neural network. To each numerical input, a weight is ascribed; through summation of the product of each input and ascribed weight, the node produces an output value. An expansion of the single-layer perceptron, multi-layer perceptron (MLP) networks composed of multiple layers of nodes linked by interconnected weights. Weights are refined through training upon labelled data, through which the input data may be classified to an output. The Recurrent Neural Network (RNN) may be considered a development of the MLP wherein output weights of intermediate nodes are also fed backwards towards the input of preceeding nodes. The inclusion of such recurrent connections renders the RNN well suited to use with temporal data, through interpretation of input datapoints with regards to their sequential context [26]. Through the gating of recurrent connections, insignificant temporal associations may be partially or entirely disregarded; a number of such implementations have been devised, varying in complexity. Discussing two of these, Chung et al. summarise functional distinctions between Gated Recurrent Units (GRU) and Long-Short Term Memory (LSTM) units. The authors noted the enhanced capability of the latter to control the amount of memory-content stored and output, through incorporation of an additional third gate [27].

Discussing applications of one-dimensional Convolutional Neural Networks (CNN) towards time-series classification, Fawaz et al. [28] liken these to the application of a sliding filter over a time-series. Kiranyaz et al. [29] discuss applications of these towards time-series classification tasks including the detection of cardiac arrhythmia and abnormal structural vibration; the authors conclude that "...even a low-power mobile device [...] will suffice to make real-time monitoring and analysis possible". Chen et al. [30] assessed the incorporation of one-dimensional CNNs within a Multi-Input Deep Convolutional Neural Network (MI-DCNN), describing this as composed of multiple parallel CNNs fed forwards towards an MLP via concatenation. The authors reported higher classification accuracies compared to conventional machine-learning methods for the classification of Arousal and Valence based upon a number of datasets consisting of multi-modal bioinformatic data.

III. Method

A. Data Capture

A multi-modal dataset was collected comprising synchronous gestural and audio recordings. For this purpose, a recording application was developed for the Apple Watch Series 8 (model: A2770) for the logging of IMU and Audio data; this was based upon Logger7 by GitHub user Shakshi3104[1]. Audio was recorded at a sample rate of 44.1 kHz through use of the built-in microphone. In addition to three-dimensional accelerometer and gyroscopic data, derived Euler angles and 4-unit quaternions are also logged. While the availability of these parallels that of the Myo, all IMU data types are logged at a higher sample rate of approximately 100 Hz; analysis of the IMU recordings indicated a mean inter-sample period of 10.075ms, with a standard deviation of 2.206ms.

Given the disparate respective sample rates of recorded IMU and Audio data, extensive time-stamping is required for the purposes of time alignment. An initial timestamp is taken as audio recording commences and a second is taken upon termination; during analyses, individual audio timestamps were interpolated between these. IMU samples are timestamped individually upon receipt.

Six violinists were recorded playing G and D major scales, two octaves in extent. These scales were intended to capture a range of both the violin's regular performance register and movement along the four strings. Participants were asked to play each note twice, capturing both an up-bow and a down-bow on each note; each bow stroke was one beat in length, at a tempo of 110BPM. Each scale was performed in two bow articulation techniques: *spiccato* and *legato*. Three takes of each exercise were recorded.

Participants comprised of undergraduate and postgraduate students and alumni of the Royal Birmingham Conservatoire.

B. Data Processing

Multimodal recordings were processed similarly to the data processing methodology detailed in [1]; this began with time-alignment and trimming. A pair of low-pass filtered RMS envelopes were used alongside a calculated threshold of 0.6x the mean audio RMS to gate concurrent audio and IMU signals, removing unwanted noise from the start and end of recordings. Audio signals were then normalised, such that their peak amplitude was equal to 1.0.

A linear de-trend function was applied to the IMU data to counteract drift. IMU data was then normalised proportionally, such that the maximum magnitude of a signal was bounded by 1, while the proportional difference in maximum magnitude

[1]https://github.com/Shakshi3104/Logger7

between concurrent channels of data (e.g. IMU Accelerometer signals X, Y, Z) was maintained. An example of a multimodal recording, processed as described, is depicted in **Figure 1**

An onset detector[2] was used as part of a system to segment audio signals into a sequence of segments representative of individual bow strokes; multi-variate IMU segments were identified whose duration coincided with the audio inter-onset-intervals produced by the onset detector. This resulted in a total of 3455 time-series segments, each comprising both audio and IMU data. Isolated segments averaged 0.47s in duration, with a standard deviation of 0.9s; this compares to an expected average duration of 0.54s at 110BPM. An average of 575.8 bow strokes wereidentified per participant. The total duration of recorded audio segments used in classification totalled 27 minutes and 10 seconds.

Sequential arrays of 13 MFCCs, 13 Delta-MFCCs, 13 Delta-Delta-MFCCs, 12 and Chroma coefficients were calculated from each individual audio segment; combined, these features intended to depict both timbral and pitch characteristics of each bow-stroke temporally. A Hanning window, 2048 samples in length, was used in computation of these, alongside a hop length of 256 samples.

Sequential IMU data and calculated MIR features were concatenated separately, into individual 3-Dimensional arrays; these were zero-padded to the length of the longest segment.

C. Data Classification

A range of DNN architectures was used for the classification of corresponding audio and gestural time-series; these included a MI-DCNN in addition to a number of further multi-input networks.

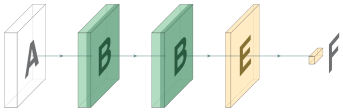

Fig. 3. Single Input DNN comprising two subnetworks

Where unimodal, input data types were classified through the use of conventional sequential DNNs; these comprised of a single input layer **(3A)**, two hidden-layers **(3B)** of types denoted by the *Network Architecture* column in **Table I**, followed by a densely connected layer **(3E)** and an output layer **(3F)**.

Multi-modal classification was conducted through the use of MI-DNNs as depicted in **Figure 4**. These comprised of two subnetworks similar in form to the aforementioned single-input networks; the final dense layers of these subnetworks were flattened **(4C)**, concatenated **(4D)**, and fed through a further fully-connected layer **(4E)** to an output layer **(4F)**. These multi-input architectures facilitate classification of the input-data modalities within a single, unified network, despite being dimensionally different; a result of disparate sample rates.

[2]https://github.com/CPJKU/madmom

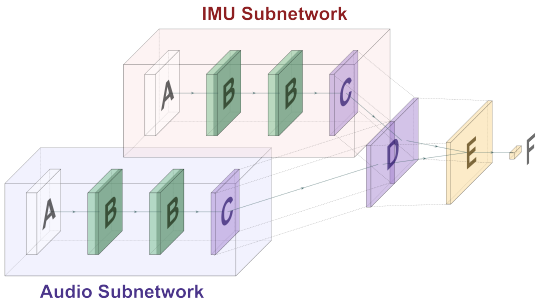

Fig. 4. Multi Input DNN comprising two subnetworks

Stratified 5-fold cross-validation was used in each instance, alongside an 80:10:10 train-test-validation split. An early stopping function was used, triggering cessation of training upon failure to reduce validation loss over 8-epochs; model weights were then restored to those which produced the lowest observed validation loss. Three classification tasks were completed: participant identification, articulation recognition, and scale recognition. The former aimed to classify input-data by participant, while the articulation recognition task aimed to classify data segments of all participants as depicting either a *spiccato* or *legato* bow stroke. The final classification task aimed to identify the scale condition from which the input-data segment was derived (D Major, G Major).

IV. Results

Test classification metrics are depicted in **Table I**, by DNN architecture and input datatype. These include test classification accuracies, calculated Receiver Operating Characteristic - Area Under Curve (ROC-AUC) values, and F-Scores. The inclusion of an additional data modality was not found to increase classification accuracies reliably across all tasks. Instead, for the identification of participants, the single-input LSTM trained upon lone IMU data proved most accurate. In the two binary classification tasks multi-input architectures achieved consistently higher classification metrics. Articulation classification accuracies were consistently high across all tested network architectures and data type combinations, with a mean classification accuracy of 99.37% achieved. Use of an MI-LSTM proved to classify input-data by scale most effectively, with an accuracy of 91.81% and F-Score of 0.918 achieved. A cross-architecture mean increase of 15.45% was observed in Scale Recognition accuracy through use of MI-DNNs trained upon multi-modal data.

In the articulation condition, p-values calculated through the use of a two-tailed t-test indicate no statistically significant differences between accuracies of the IMU and Audio single-input classifiers ($p = 0.712$). Likewise, differences in classification accuracy between multi- and single-input classification accuracies exhibited no statistical significance ($p = 0.473$), although a lesser p-value is observed.

Towards participant classification, the higher classification accuracies observed through use of IMU data, versus audio, proved significant at $p < 0.01$. Single-input IMU classifiers

TABLE I
CLASSIFICATION METRICS BY DATA TYPE, TASK, AND DNN ARCHITECTURE

Data Type	Network Architecture		Participant Recognition			Articulation Recognition			Scale Recognition		
			Acc (%)	AUC	F Score	Acc (%)	AUC	F Score	Acc (%)	AUC	F Score
Audio	MLP		76.38	0.948	0.765	99.29	0.999	0.993	71.44	0.803	0.715
”	LSTM		82.43	0.967	0.828	99.51	0.999	0.995	72.00	0.822	0.721
”	1D-CNN		80.75	0.964	0.812	99.43	0.999	0.994	73.41	0.839	0.737
”	GRU		79.39	0.957	0.796	99.16	0.999	0.990	70.65	0.810	0.710
IMU	MLP		94.41	0.991	0.947	99.05	0.998	0.990	72.41	0.796	0.718
”	LSTM		96.43	0.996	0.965	99.51	0.990	0.994	72.63	0.788	0.720
”	1D-CNN		96.00	0.994	0.961	99.35	0.997	0.992	79.12	0.868	0.788
”	GRU		91.83	0.989	0.918	99.29	0.993	0.992	63.83	0.698	0.630
Audio†+ IMU	MLP†	MLP	94.66	0.993	0.950	99.67	0.999	0.997	84.22	0.923	0.841
”	LSTM†	LSTM	93.85	0.988	0.938	99.56	0.998	0.996	91.81	0.964	0.918
”	1D-CNN†	1D-CNN	94.20	0.986	0.942	99.67	0.999	0.997	88.40	0.935	0.885
”	GRU†	GRU	95.59	0.995	0.956	98.49	0.989	0.985	88.45	0.951	0.885
”	MLP†	1D-CNN	94.08	0.992	0.942	99.56	0.999	0.996	86.55	0.935	0.866
”	1D-CNN†	MLP	93.68	0.992	0.937	99.62	0.999	0.996	83.68	0.917	0.837
”	LSTM†	1D-CNN	93.79	0.987	0.936	99.56	0.998	0.996	88.70	0.939	0.887
”	1D-CNN†	LSTM	91.24	0.979	0.911	99.35	0.997	0.993	87.31	0.937	0.873

achieved a marginally higher mean-accuracy rate of 94.6675 (versus 93.88625), although this increase did not prove statistically significant ($p = 0.426$).

In the Scale classification condition, no statistically significant difference was observed between the classification accuracies of the Single-input IMU and Audio classifiers ($p = 0.971$). MI-DNN implementations, however, were observed to increase classification accuracy, from a mean of 71.93 across single-input DNNs to 87.390. This result proved statistically significant, ($p < .00001$), indicating that task performance can - in some instances - be enhanced through inference in multiple modalities, beyond the capabilities of individual networks trained upon singular modalities.

V. DISCUSSION

The outlined results demonstrate the significance of architecture selection towards tasks undertaken; significant variation was observed per-task despite no change to the processing of input data. Further study may be required for the evaluation of best practice data processing techniques for use in such implementations. These may include additional or alternate MIR features, and feature derivation of the IMU data.

Lone datatype classification accuracies were not a reliable predictor of network-architecture performance following incorporation within a MI-DNN. Intuitively, one might expect that combining the two single-input models which performed best for any given task into a MI-DNN would prove most effective - or at least as accurate as the most accurate subnetwork, given the theoretical ability of post-concatenation dense layers to discriminate between useful and non-useful node inputs via weighting. This was not consistently observed, however.

In the participant classification condition, while MI-DNN train accuracies approached 1.00, validation and test accuracies failed to do so – to a greater extent than was observed during the training of single-input networks upon IMU data. Marginally higher average train accuracies, combined with marginally lower average validation accuracies, suggest overfitting. These can be observed in **Figure 5**, wherein bold stepped lines represent mean validation accuracies per epoch, shaded regions indicate the range between the maximum and minimum validation accuracies per epoch, dashed stepped lines depict mean train accuracies per epoch, and hatched regions indicate the range between the maximum and minimum train accuracies per epoch. Minimum, maximum, and mean accuracies depicted here are averaged across network architectures. Observed overfitting may have been caused by a resultant network with a greater number of parameters training upon a relatively small dataset. Collation of a larger dataset may facilitate an assessment of the cause of the observed drop in accuracy. While a cause cannot be confirmed from these results alone, MI-DNN architectures should not be assumed to outperform any one component subnetwork; as observed in the scale-recognition task, however, multi-input networks may prove capable of outperforming each of their component subnetworks in some instances.

MI-DNNs achieved greater participant classification accuracies than single-input networks trained solely upon audio features; in comparison, these exhibited significant underfitting, with train and validation accuracies diverging both at a faster rate and to a greater degree.

Overall, participant classification accuracies indicated that both gestural and audible distinctions between individual violinists proved identifiable, albeit to varying extents. Single-input audio networks achieved accuracies comparable to those reported by D'Amato et al. [15], while both multi- and single-input networks trained upon recorded IMU data exceeded these by around 14 percentage points - albeit through the use of

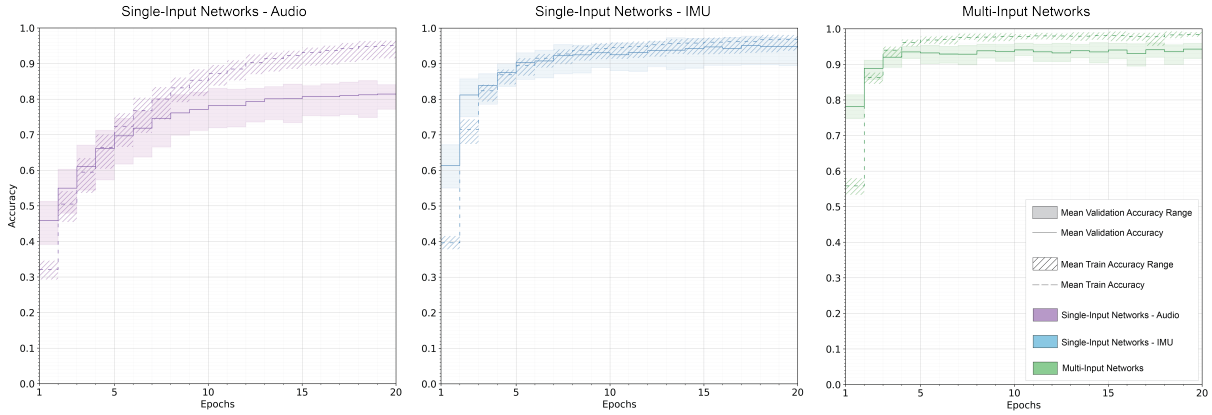

Fig. 5. Participant Classification - Mean Validation and Training Accuracies per epoch

disparate recording technologies and classification algorithms.

Trained networks were able to distinguish *spiccato* and *legato* bow strokes reliably across participants, robust to performer idiosyncracies. Presented articulation classification accuracies indicate a similar level of accuracy to those observed in prior studies [7], [9], [13], although in a notably binary implementation; these cited works detail the classification of 4, 3, and 7 bow articulation techniques, respectively. A demonstrated ability of the trained models to recognise scale, despite only a single pitch class difference (C vs. C#), would suggest that some quantitive, note-to-note distinction manifests as a result of situation within a different sequence.

VI. Conclusion

The results presented indicate the potential suitability of a ubiquitous approach towards the conduction of multimodal musicological analyses. While MI-DNN architectures did not consistently improve classification accuracies beyond those of single-input architectures, the instance in which they did so proved statistically significant; on balance, the incorporation of multimodality in such tasks may improve classification accuracies, but should not be assumed to do so. Due to computational expense, the utility of smartwatch-based DNN implementation will likely remain limited towards offline applications, and thus may prove to be of limited use in a performance environment. Instead, such approaches may be better suited towards applications such as violin practice-feedback. The emergence of commercially-accessible wearable sensors has facilitated convenience in multimodal data capture; further consideration of such practically-accessible multimodal interfaces may prove opportune during the development of novel musical aids, offering feasible, real-world utility. Mobile devices, such as the smartwatch, may be considered comparatively ubiquitous in contrast to specialist hardware technologies used in prior implementations, facilitating the democratisation of both research conduction and the products thereof. Through wider, virtual distribution, the developed recording method may facilitate remote participation, allowing for the collation of a far larger dataset upon which multimodal analysis techniques may be refined.

VII. Acknowledgments

Funding was provided towards the completion of this research by Birmingham City University. All work was subject to ethical approval.

VIII. References

[1] W. Wilson, N. Granieri, and I. Ali-MacLachlan, "Time's up for the Myo? The smartwatch as a ubiquitous alternative for audio-gestural analyses.," en, in *Proceedings of the 23rd International Workshop on New Interfaces for Musical Expression (NIME '23)*, Mexico City, Mexico: Universidad Autónoma Metropolitana, 2023.

[2] K. C. Ng, T. Weyde, O. Larkin, K. Neubarth, T. Koerselman, and B. Ong, "3D Augmented Mirror: A Multimodal Interface for String Instrument Learning and Teaching with Gesture Support," en, in *Proceedings of the 9th international conference on Multimodal interfaces*, Nagoya Aichi Japan: ACM, Nov. 2007, pp. 339–345, ISBN: 978-1-59593-817-6. DOI: 10.1145/1322192.1322252.

[3] V. D'Amato, E. Volta, L. Oneto, G. Volpe, A. Camurri, and D. Anguita, "Accuracy and Intrusiveness in Data-Driven Violin Players Skill Levels Prediction: MOCAP Against MYO Against KINECT," en, in *Advances in Computational Intelligence: 16th International Work-Conference on Artificial Neural Networks, IWANN 2021, Virtual Event, June 16–18, 2021, Proceedings, Part II*, ser. Lecture Notes in Computer Science, vol. 12862, Cham: Springer International Publishing, 2021, pp. 367–379, ISBN: 978-3-030-85098-2 978-3-030-85099-9. DOI: 10.1007/978-3-030-85099-9.

[4] D. Dalmazzo and R. Ramirez, "Air violin: A Machine Learning Approach to Fingering Gesture Recognition," en, in *Proceedings of the 1st ACM International Workshop on Multimodal Interaction for Education*, Glasgow UK: ACM, Nov. 2017, pp. 63–66, ISBN: 978-1-4503-5557-5. DOI: 10.1145/3139513.3139526.

[5] Y. Wang, B. Zhang, and O. Schleusing, "Educational violin transcription by fusing multimedia streams," en, in *Proceedings of the international workshop on Educational multimedia and multimedia education*, Augsburg Bavaria Germany: ACM, Sep. 2007, pp. 57–66, ISBN: 978-1-59593-783-4. DOI: 10.1145/1290144.1290154.

[6] G. Volpe, K. Kolykhalova, E. Volta, *et al.*, "A multimodal corpus for technology-enhanced learning of violin playing," en, in *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter*, Cagliari Italy: ACM, Sep. 2017, pp. 1–5, ISBN: 978-1-4503-5237-6. DOI: 10.1145/3125571.3125588.

[7] D. Dalmazzo, G. Waddell, and R. Ramírez, "Applying Deep Learning Techniques to Estimate Patterns of Musical Gesture," en, *Frontiers in Psychology*, vol. 11, p. 575 971, Jan. 2021, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2020.575971.

[8] D. Dalmazzo, S. Tassani, and R. Ramírez, "A Machine Learning Approach to Violin Bow Technique Classification: A Comparison Between IMU and MOCAP systems," en, in *Proceedings of the 5th international Workshop on Sensor-based Activity Recognition and Interaction*, Berlin Germany: ACM, Sep. 2018, pp. 1–8, ISBN: 978-1-4503-6487-4. DOI: 10.1145/3266157.3266216.

[9] Á. Sarasúa, B. Caramiaux, A. Tanaka, and M. Ortiz, "Datasets for the Analysis of Expressive Musical Gestures," en, in *Proceedings of the 4th International Conference on Movement Computing*, London United Kingdom: ACM, 2017, pp. 1–4, ISBN: 978-1-4503-5209-3. DOI: 10.1145/3077981.3078032.

[10] A. Lucas, F. Schroeder, and M. Ortiz, "Enabling Communities of Practice Surrounding the Design and Use of Custom Accessible Music Technology," en, *Computer Music Journal*, vol. 44, no. 2-3, p. 17, Jul. 2020, ISSN: 0148-9267, 1531-5169. DOI: 10.1162/comj_a_00567.

[11] F. Morreale and A. P. McPherson, "Design for Longevity: Ongoing Use of Instruments from NIME 2010-14," en, in *Proceedings of the International Conference on New Interfaces For Musical Expression*, Copenhagen, Denmark: Aalborg University, May 2017, pp. 192–197.

[12] J. J. Craig, *Introduction to Robotics - Mechanics and Control*, en, 3rd. New Jersey, USA: Pearson Education, Inc., 2005, ISBN: 0-13-123629-6.

[13] D. Dalmazzo and R. Ramirez, "Bowing Gestures Classification in Violin Performance: A Machine Learning Approach," *Frontiers in Psychology*, vol. 10, p. 344, Mar. 2019. DOI: 10.3389/fpsyg.2019.00344.

[14] C. Auepanwiriyakul, S. Waibel, J. Songa, P. Bentley, and A. A. Faisal, "Accuracy and Acceptability of Wearable Motion Tracking for Inpatient Monitoring Using Smartwatches," en, *Sensors*, vol. 20, no. 24, p. 7313, Dec. 2020, ISSN: 1424-8220. DOI: 10.3390/s20247313.

[15] V. D'Amato, E. Volta, L. Oneto, G. Volpe, A. Camurri, and D. Anguita, "Understanding Violin Players' Skill Level Based on Motion Capture: A Data-Driven Perspective," en, *Cognitive Computation*, vol. 12, no. 6, pp. 1356–1369, Nov. 2020, ISSN: 1866-9956, 1866-9964. DOI: 10.1007/s12559-020-09768-8.

[16] M. Schedl, E. Gómez, and J. Urbano, "Music Information Retrieval: Recent Developments and Applications," en, *Foundations and Trends in Information Retrieval*, vol. 8, no. 2-3, pp. 127–261, 2014, ISSN: 1554-0669, 1554-0677. DOI: 10.1561/1500000042.

[17] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," en, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980, ISSN: 0096-3518. DOI: 10.1109/TASSP.1980.1163420.

[18] Y. Wu, Q. Wang, and R. Liu, "Music Instrument Classification using Nontonal MFCC," en, in *Proceedings of the 2017 5th International Conference on Frontiers of Manufacturing Science and Measuring Technology (FMSMT 2017)*, Taiyuan, China: Atlantis Press, 2017, ISBN: 978-94-6252-331-9. DOI: 10.2991/fmsmt-17.2017.88.

[19] T. Li and M. Ogihara, "Music Artist Style Identification by Semi-Supervised Learning from both Lyrics and Content," in *Proceedings of the 12th ACM International Conference on Multimedia*, Jan. 2004, pp. 364–367. DOI: 10.1145/1027527.1027612.

[20] M. I. Mandel and D. P. W. Ellis, "Song-Level Features and Support Vector Machines for Music Classification," en, in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, London, UK, 2005, p. 6.

[21] H. S. Alar, R. O. Mamaril, L. P. Villegas, and J. R. D. Cabarrubias, "Audio classification of violin bowing techniques: An aid for beginners," en, *Machine Learning with Applications*, vol. 4, p. 100 028, Jun. 2021, ISSN: 26668270. DOI: 10.1016/j.mlwa.2021.100028.

[22] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," en, *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, Nov. 2001, ISSN: 1000-9000, 1860-4749. DOI: 10.1007/BF02943243.

[23] S. S. Stevens, J. Volkmann, and E. B. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," en, *Journal of the Acoustical Society of America*, vol. 8, no. 185, 1937. DOI: 10.1121/1.1915893.

[24] B. McFee, C. Raffel, D. Liang, *et al.*, "Librosa: Audio and Music Signal Analysis in Python," en, Austin, Texas, 2015, pp. 18–24. DOI: 10.25080/Majora-7b98e3ed-003.

[25] E. Alpaydin, *Introduction to Machine Learning*, 4th. Massachusetts, USA: The MIT Press, 2020, ISBN: 978-0-262-04379-3.

[26] S. Russell and P. Norvig, *Artificial Intelligence - A Modern Approach.* (Pearson series in Artificial Intelligence),

4th edn. New Jersey, USA: Pearson Education, Inc., 2020, ISBN: 978-0-13-461099-3.

[27] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," en, in *NIPS 2014 Workshop on Deep Learning*, Montreal, Canada, Dec. 2014.

[28] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: A review," en, *Data Mining and Knowledge Discovery*, vol. 33, no. 4, pp. 917–963, Jul. 2019, ISSN: 1384-5810, 1573-756X. DOI: 10.1007/s10618-019-00619-1.

[29] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," en, *Mechanical Systems and Signal Processing*, vol. 151, p. 107 398, Apr. 2021, ISSN: 08883270. DOI: 10.1016/j.ymssp.2020.107398.

[30] P. Chen, B. Zou, A. N. Belkacem, *et al.*, "An improved multi-input deep convolutional neural network for automatic emotion recognition," en, *Frontiers in Neuroscience*, vol. 16, p. 965 871, Oct. 2022, ISSN: 1662-453X. DOI: 10.3389/fnins.2022.965871.