Ensemble of Multimodal Deep Learning Models for Violin Bowing Techniques Classification

Zain Muhammed [1], Nagamanoj Karunakaran [2,*], Pranamya P. Bhat [1], and Arti Arya [1]

[1] Department of Computer Science Engineering, PES University, Bengaluru, India
[2] Control Design Automation, MathWorks, Bengaluru, India
Email: zainmuhammed66@gmail.com (Z.M.); nkarunak@mathworks.com (N.K.); pranamyabhat27@gmail.com (P.P.B.); artiarya@pes.edu (A.A.)
*Corresponding author

***Abstract*—Bowing gesture while playing violin refers to the motion of the violinist's arm. Violinists use different types of bow strokes to express musical phrases, played by the movement of the right arm holding the fiddle bow. Although the sound produced by each bow stroke is distinct, it can be difficult for new fiddlers to distinguish and recognize these bowing techniques. So, this paper presents a novel approach of an ensemble of multimodal deep learning models consisting of one Convolution Neural Network (CNN) and two Long Short-Term Memory (LSTM) models to classify into one of the five bowing classes: detaché, legato, martelé, spiccato and staccato. The dataset used consists of audio samples performed by 8 violinists along with the motion of their forearms measured using a Myo sensor device, to acquire 8-channels of Electromyogram (EMG) data and 13-channels of Inertial Measurement Unit (IMU) data. The audio features are extracted from audio excerpts and time domain features are extracted from EMG and IMU motion signals. These features are passed into an ensemble of deep learning models to make the final prediction using weighted voting. The proposed ensemble classifier was able to deliver optimal results with an overall accuracy of 99.5%, which is better than the previous studies that took only either audio or motion data into consideration.**

***Keywords*—violin bowing technique, audio features, motion features, Electromyogram (EMG), Inertial Measurement Unit (IMU), Essentia, Convolution Neural Network (CNN), Long Short-Term Memory (LSTM), deep learning model**

I. Introduction

Music performers narrate music to their audience with expression and emotion, which forms the bridge of emotional communication between them. With instrumental music, the expression is projected by producing the variation in characteristics of sound such as dynamics, phrasing, timbre and articulation to bring the music to life. In bowed string instruments such as violin, the left hand is responsible only for pitch and vibrato, and the bow (right) hand can affect the sound with only three actions, i.e., bow speed, bow pressure and the change of sounding point. These three elements correspond to the three dimensions that know in our physical world. Forward and backward action that's bow speed, up and down that's pressure and sideways that the sounding point [1]. The manner in which these elements are controlled, by the movement of the bowed (right) arm defines a bowing technique. Therefore, the sound produced and the movement on the right arm defines the type of bowing stroke, i.e., the bowing technique.

Artificial intelligence has made significant progress in the field of music. Music genre classification [2], automated music generation, music recommendation system and so on depicts the blend of technology and music. However, the exploration on the grounds of stringed instruments such as violin, sarangi, and cello which been limited. Detection of these gestures in a violin sample requires expertise in the field of music. Hence, beginner violinists find quite a challenge to distinguish between the bowing techniques. In order to address this, the paper proposes the usage of deep learning methods in the identification of bowing gestures. The five commonly used bowing techniques have been chosen for the study, namely: detaché, legato, martelé, spiccato, staccato.

The dataset considered [3] includes audio excerpts played by eight different participants using the five unique bowing techniques. The Myo device attached to the participant's bow arm provides accurate readings of the rotational force and the directional attributes. These readings recorded assist in capturing the gestural expressions in the musical content. Both audio features, as well as motion features are taken into consideration for the analysis.

Traditional machine learning algorithms were applied for bowing technique classification using either audio features or motion features. It's seen that "detaché" is better classified in audio classifiers, whereas the motion classifier performed well in classifying "martelé". In order to address this issue, this paper proposes a novel approach of ensemble of deep learning models consisting of a Convolution Neural Network (CNN) model for classification using audio features and two Long-Short Term Memory (LSTM) models that classify using the motion features. The main contribution of this paper is using multimodal features namely audio features and

Manuscript received May 6, 2023; revised June 5, 2023; accepted July 2, 2023; published January 9, 2024.

motion features together to classify a specific bowing technique into five different classes. So far in the available literature, either of the two features is considered for classification.

The rest of the paper is organized as follows: Section II describes the literature available for bowing classification using audio and motion features and the research gaps identified. Section III highlights some background details necessary to understand different bowing techniques for a violin player. Section IV discusses about the dataset, feature extraction process and the proposed classification technique for different bowing forearm movements. Section V details about the experimental setup, experiments carried out and the results obtained. Section VI concludes the work done and challenges faced during implementation and new research directions.

II. Related Work

A. Hand Gesture Recognition Using EMG and IMU

There have been several studies to recognize human gestures using Electromyography (EMG) and Inertial Measurement Unit (IMU) acquired from the Myo armband device. Tepe and Demir [4] presented a study on detecting and classifying muscle activation in EMG data acquired using the Myo armband. This study has extracted time domain features namely Mean Average Value (MAV), Root Mean Square (RMS), Waveform Length (WL), Slope Sign Change (SSC) and Zero Crossing (ZC) from EMG motion signal and used Support Vector Machine (SVM) classifier and got an average accuracy 98.75%. Tepe and Demir [5] presented the classification of hand gestures from EMG Myo armband data using SVM. In this study, time domain as well as frequency domain features were considered to get an average accuracy of 95.83%. Georgi *et al.* [6] have presented recognition of hand and finger gestures with EMG-based muscle activity sensing and IMU-based motion separately having obtained an average accuracy of 92.8% with IMU sensors and 85.1% with EMG sensors. This study showed that EMG as well as IMU systems can be used for hand gesture recognition. Ali and Yanen [7] have presented the SVM classification of controls of 6 DOF-robot using fused IMU and EMG features. With IMU features such as RMS and MAV and EMG features like Myopulse Percentage Rate and Average Amplitude Change, this study got an accuracy of above 98%. Dezhen *et al*. [8] studied and provided a comprehensive review of deep learning in EMG pattern recognition for human-machine interfaces. This study summarizes the opportunities, advantages, and challenges to use deep learning in solving questions involving EMG recognition. Vasconez *et al*. [9] have presented hand gesture recognition using both EMG IMU signals and Deep Q-Network (DQN) reinforcement learning algorithm and have obtained an accuracy of 97.5%, 80.04%, 84.49% using DQN, K-Nearest Neighbors (KNN) and CNN classifiers, respectively.

B. Bowing Technique Classification

Mukherjee and Anand [10] presented the classification of cello bowing using motion features extracted from 3-axis accelerometer, gyroscope and magnetometer gathered from the wearable Orient wireless sensors—one attached under the frog of the bow and the other on the wrist of the playing hand. This study has used Support Vector Machine (SVM) to classify bowing techniques such as legato, staccato, martelé, spiccato, tremolo, col legno & ricochet with 95% accuracy. Dalmazzo and Ramirez [11] presented bowing technique classification by applying Hierarchical Markov Model (HHMM) to data from inertial sensors and audio recordings acquired from a single violinist playing a simple G-Major scale and the accuracy obtained with motion only, audio only and audio + motion features are 93.2%, 39.01% and 94.61%, respectively. Sun *et al*. [12] presented the use of Deep learning models for classifying violin bowing techniques by analyzing the signals from inertial sensors and depth camera and were able to get an average accuracy greater than 80%. Dalmazzo and Ramirez [13] presented Deep Learning techniques for classifying violin bowing techniques such as detaché, legato, martelé, collé, staccato, ricochet, trémolo and col legno and were able to get an accuracy of 97.15%, 98.55%, 99.23% with CNN, 3D-MultiHeaded CNN, CNN LSTM models respectively. This study has used a dataset that has only IMU features extracted from recordings of simple G-Major scale. Hernan *et al*. [14] presented an application of CNN to classify the bowing techniques such as detaché, legato, ricochet, spiccato and double-stops using audio features with an accuracy of 94.8%. Alvaro *et al*. [3] classified the bowing technique such as detaché, martelé, legato, staccato and spiccato using Hidden Markov Model and Sequential Monte Carlo Model and were able to get an average accuracy of 94.3%, 41.2%, 40.7% on the three different datasets with IMG + EMU features.

The previous studies have presented methods that incorporate either motion data [11–14] or audio data [10] and dataset considered was from a simple playing of violin such as scale. To the best of our knowledge, there is no work available where both motion and audio features are considered for classifying bowing categories. In the proposed approach, both audio and motion features extracted from excerpts of violinists playing Kreutzer Study No. 2 in C-Major [3] are considered for effective classification.

III. Musical Materials

Bowing allows musicians to express a wide variety of emotions. The pressure exerted on the bow, the position of the bow on the strings, the distribution of the bow's weight and the inclination of the bow are all factors that impact the sound produced by the violin. Different combinations of these factors constitute the different bowing gestures. The following bowing techniques have been considered [3, 11]:

- ***Detaché***: *Meaning separated*. The most fundamental bowing technique in the violin repertoire. It involves a constant bow speed as

well as constant bow pressure, while moving smoothly from one note to note keeping the sound dynamically stable.

- ***Martelé***: *Meaning hammered*. It is a form of detaché with a more distinct attack. The motion starting point is given emphasis by using a faster and slightly stronger initial movement. At the end of the movement, there is a moment of silence.
- ***Legato***: A technique used to play more than one note on the same bow stroke. Excessive accents, attacks or emphasis are avoided by the musician.
- ***Spiccato***: The slowest of bowing strokes. The relaxed manner in which the bow is held as well as its bouncing results in a series of short, distinct notes. The strings are attacked on a vertical angular approach of the bow with a controlled weight and precise hand-wrist control.
- ***Staccato***: A variation of martelé which is shorter and sharper. Controlled pressure over the string followed by an accentuated release in the direction of the bow-stroke results in staccato having a clean attack.

The dataset containing these five bowing techniques and their classification is discussed in Section IV.

IV. Methodology

A. Dataset

The dataset used in this paper was obtained from Álvaro *et al*. [3]. All the bowing techniques were established in Kreutzer's Study No. 2 in C-Major, a standard pedagogic repertoire. The dataset includes the multimodal (audio and motion) recordings of eight professional violinists performing the repertoire, 10 times each, using the following bowing techniques: legato, detaché, martelé, staccato and spiccato, this is called as *primary dataset*. The second set of data includes the recordings of the same violinist playing the same repertoire, 10 times each, with only three bowing techniques legato, detaché and spiccato, varying the dynamics from pianissimo to fortissimo (soft to loud), this is called as *dynamic dataset*. The third set of data includes the recordings of violinists playing the repertoire, 10 times, with the same three bowing techniques but accelerating the speed from slow to fast, this is called as *tempo dataset*. Samples from all the three datasets *primary*, *dynamic* and *tempo* are put together and this is called *full dataset*. Table I shows the distribution of the samples for each bowing technique across the three different datasets. The dataset captures audio samples recorded at a sample rate of 44100 Hz along with the motion data of the forearm of each participant. The dataset includes the motion data, 8-channels of electromyography (EMG) data in a circular formation around the right forearm of each participant acquired from the Myo device and IMU data consisting of 3-axis gyroscope (x,y,z), 3-axis accelerometer (x,y,z) and the Myo's calculated position data which provides Euler angles for pitch, roll and yaw along with Unit quaternions.

TABLE I. Distribution of Samples

Bowing Techniques	Primary Dataset	Dynamic Dataset	Tempo Dataset
detaché	80	80	80
legato	80	80	80
spiccato	80	80	80
martelé	80	-	-
staccato	80	-	-

B. Feature Extraction

1) Audio feature extraction: For each audio sample around 73 features [2] were extracted using Essentia [15], an open-source C++ library for audio analysis and audio-based music information retrieval. Segmentation of the input audio signal was performed by creating multiple frames of equal size along with a constant hop size to hop in between frames. Features were then extracted for each frame and their mean was taken. Fig. 1 represents the audio input signal for the excerpt played using legato along with its features in the form of a spectrogram. Essentia provides a large set of spectral, temporal, high-level and tonal descriptors. The audio features extracted belong to the following categories:

- **Spectral features**: These features provide general frequency-domain metrics of the audio signal. It also describes the shape of the spectrum. E.g.: Mel-Frequency Cepstral Coefficients (MFCCs), Gammatone Frequency Cepstral Coefficients (GFCCs), Flux, High Frequency Content (HFC).
- **Temporal features**: These include time domain features of the audio signal. E.g.: RMS, Zero crossing rate, etc.
- **Tonal features**: These features describe the arrangement of chords and keys in the audio signal. Examples include chord descriptors, pitch salience, dissonance, etc.

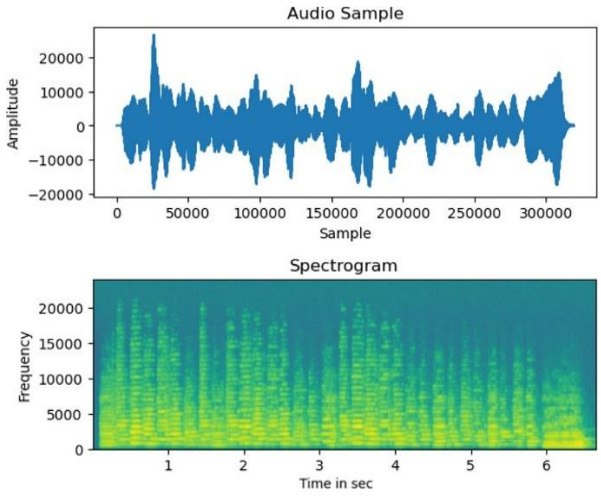


Fig. 1. Input audio signal along with its spectrogram.

2) Motion feature extraction: The motion data consists of 8 channels of EMG data along with IMU data that contains unit quaternions, Euler angles for pitch, yaw, roll and 3-axis accelerometer, 3-axis gyroscope

adding up to a total of 21 input signals for each recording. For each of the input signals, five time-domain features were captured [16–20]. These include: Mean Absolute Value (MAV), Root Mean Square (RMS), Waveform Length (WL), Zero Crossing (ZC) and Slope Sign Change (SSC) [16]. The input signal was segmented into windows of length 1s along with an overlap of 50%. The features were individually calculated for each window [16]. RMS determines the power of the signal in the time domain whereas SSC and ZC determine a measure of the frequency content of the input signal [16]. Fig. 2 represent the input signal for one channel of EMG data taken for one recording along with the RMS and MAV computed for each window where definitions of RMS, MAV, WL, ZC and SSC are given in Eqs. (1)–(5), respectively.

$$RMS = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(x_i)^2} \quad (1)$$

$$MAV = \frac{1}{n}\sum_{i=1}^{n}|x_i| \quad (2)$$

$$WL = \frac{1}{n}\sum_{i=1}^{n-1}|x_{i+1} - x_i| \quad (3)$$

$$ZC = \sum_{i=1}^{n} sgn(-x_i x_{i+1}) \quad (4)$$

$$sgn(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

$$SSC = \sum_{i=2}^{n} f[(x_i - x_{i-1}) \times (x_i - x_{i+1})] \quad (5)$$

$$f(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

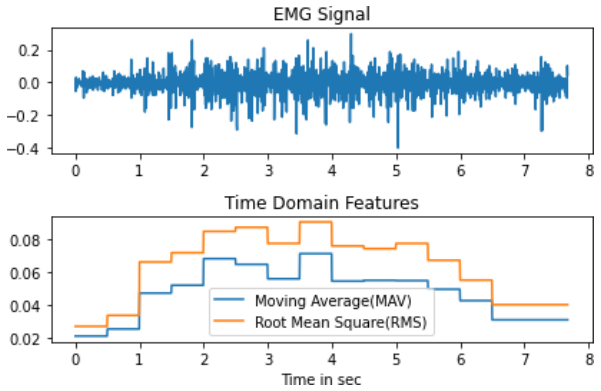


Fig. 2. EMG for one channel combined with RMS, MAV.

C. Classification of Violin Bowing Technique Using ML Algorithms

Following feature extraction, around 73 audio features and 105 motion features were obtained. The dataset is divided into four parts: original dataset containing all samples (880), the initial samples played using all five bowing techniques (400), the second variation wherein the dynamics were varied (240) and the third variation wherein the tempo was varied (240). To prevent overfitting, dimensionality reduction was performed on the audio as well as motion features using Principal Component Analysis (PCA). Fig. 3 represents a graph indicating the change in variance vs. the number of principal components for audio features. From the graph (Fig. 3), we can see that approximately 97% of the variance is explained when around 40 principal components are considered. A similar experiment was performed for motion features as well and it was observed that 30 principal components were required to explain approximately 97% of the variance.

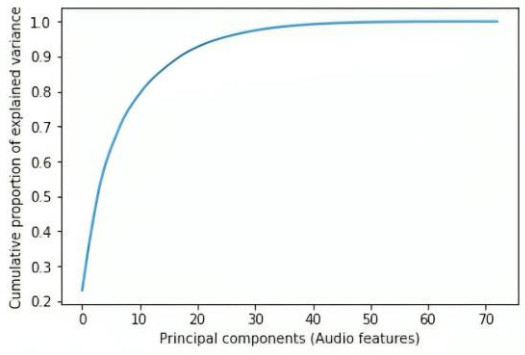
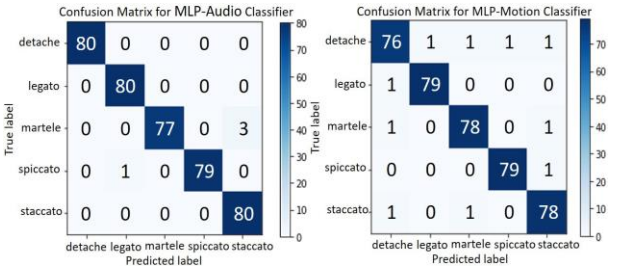


Fig. 3. PCA analysis for audio features.

After selecting 40 and 30 principal components for the audio and motion features respectively, multiple machine learning models were trained and tested on all four datasets for audio features and motion features separately. 10-fold cross-validation was used to validate the models. The following ML models were used: Logistic Regression, K-Nearest Neighbor, SVM (linear), Naive Bayes model and Multi-layer Perceptron (MLP). Table II presents the accuracy of each model on all four datasets for audio features as well as motion features individually.

TABLE II. ML MODEL ACCURACIES ON ALL DATASETS

ML Models	Full		Primary		Dynamic		Tempo	
	Audio	Motion	Audio	Motion	Audio	Motion	Audio	Motion
LR	99.3	90.2	98.8	92	100	97.5	100	97.1
KNN	98.4	98.4	97.8	99.5	100	99.2	100	95
SVM	99.9	88.2	97	92.7	100	97.1	100	98.8
NB	82.7	81.7	85.5	81	92.1	89.2	95.8	86.7
MLP	99.7	97.6	99	97.5	100	98.8	100	99.2



(a) (b)

Fig. 4. Confusion matrices of MLP with (a) audio and (b) motion feature.

From Table II, it is observed that audio features have a greater impact as compared to motion features. The Naïve Bayes classifier gives the least accuracy whereas the MLP classifier gives the highest accuracy for all datasets. From Fig. 4, it can be seen that "detaché" is better classified in audio classifiers, "martelé" is better classified in motion classifiers. In order to address the need for both audio and motion features for better

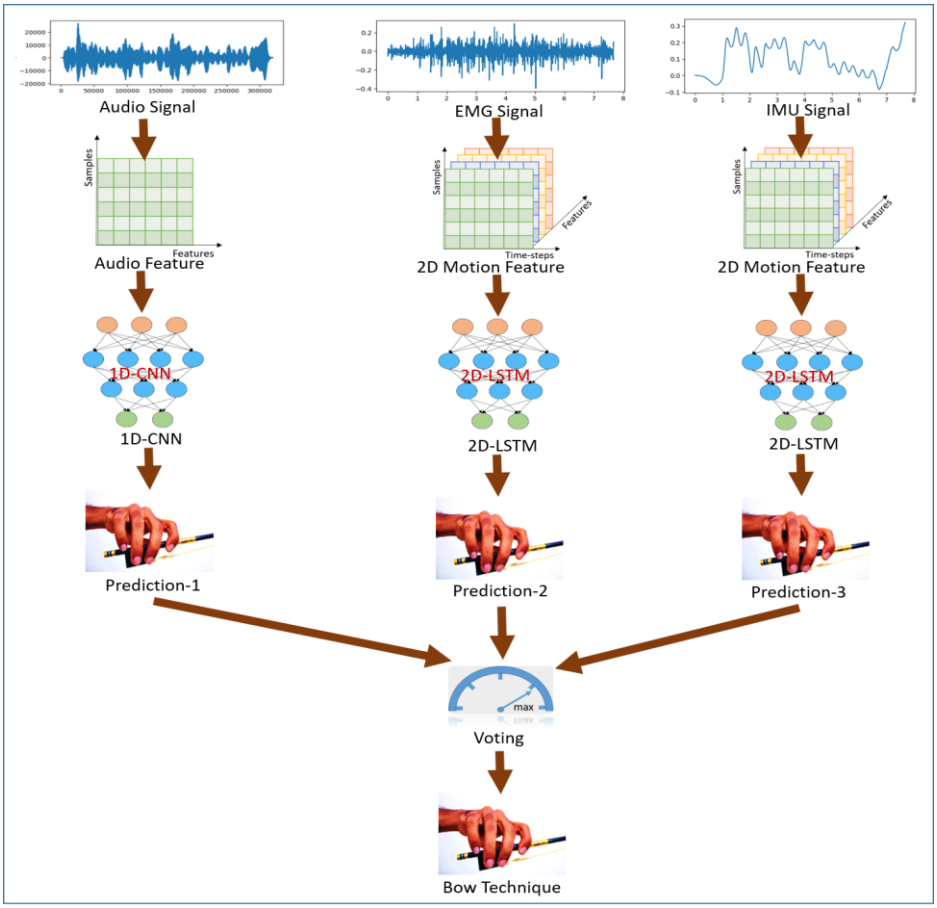


Fig. 5. System diagram of proposed ensemble of deep learning models for bowing technique classification.

classification of violin bowing techniques, this paper proposes a novel approach of ensemble of deep learning models consisting of a CNN model for classification using audio features and two LSTM models that classify using the motion features.

D. *Classification of Violin Bowing Technique Using Ensemble of Multimodal Deep Learning Models*

The system diagram of the proposed ensemble of deep learning models comprising of a CNN model along with two LSTM models is as shown in Fig. 5.

Fig. 6 shows the architecture of the CNN model, consisting of two convolution layers and an output layer, that predominantly uses the audio features extracted from the violin excerpt. The features are extracted from Essentia and PCA is applied to select 67 features for (N)samples. The dataset forming the input shape of [N, 67] is used to train the CNN model. For instance, with *primary dataset* we have an input shape of [400, 67]. Once the model is trained using this data, it compiles the first stage of prediction with respect to the audio features.

Further, EMG signals are segmented into windows of length 0.25 s along with an overlap of 50% to extract time domain features over each window to give 40 features in 42 timesteps for (N)samples. The dataset forming the input shape of [N, 42, 40], is used to train the first 2D-LSTM model. For instance, with *primary dataset* we have an input shape of [400, 42, 40]. Fig. 7 shows the architecture for the first LSTM model consisting of an LSTM layer and a dense layer with drop out before the output layer. This contributes to the second set of predictions.

```
CNN Model Summary for Audio Features
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 64, 25)            275

 batch_normalization (BatchN  (None, 64, 25)           100
 ormalization)

 re_lu (ReLU)                (None, 64, 25)            0

 max_pooling1d (MaxPooling1D  (None, 32, 25)           0
 )

 conv1d_1 (Conv1D)           (None, 32, 68)            17068

 batch_normalization_1 (Batc  (None, 32, 68)           272
 hNormalization)

 re_lu_1 (ReLU)              (None, 32, 68)            0

 max_pooling1d_1 (MaxPooling  (None, 16, 68)           0
 1D)

 flatten (Flatten)           (None, 1088)              0

 dense (Dense)               (None, 5)                 5445

=================================================================
Total params: 23,160
Trainable params: 22,974
Non-trainable params: 186
_________________________________________________________________
```

Fig. 6. CNN architecture for audio data.

LSTM-EMG Model Summary for Motion-EMG Features
Model: "sequential_1"

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 100)	56400
dropout (Dropout)	(None, 100)	0
dense_1 (Dense)	(None, 100)	10100
dense_2 (Dense)	(None, 5)	505

Total params: 67,005
Trainable params: 67,005
Non-trainable params: 0

Fig. 7. LSTM architecture for EMG data.

LSTM-IMU Model Summary for Motion-IMU Features
Model: "sequential_2"

Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, 100)	66400
dropout_1 (Dropout)	(None, 100)	0
dense_3 (Dense)	(None, 100)	10100
dense_4 (Dense)	(None, 5)	505

Total params: 77,005
Trainable params: 77,005
Non-trainable params: 0

Fig. 8. LSTM architecture for IMU data.

Lastly, IMG signals are segmented into windows of length 0.25s along with an overlap of 50% to extract time domain features over each window to give 65 features in 42 time-steps for (N)samples. The dataset forming the input shape of [N, 42, 65], is used to train the second 2D-LSTM model. For instance, with *primary dataset* we have an input shape of [400, 42, 65]. The architecture used for the second LTSM model consisting of an LSTM layer and a dense layer with drop out before the output layer is shown in Fig. 8. This model provides the third set of predictions. As a final step, the predictions generated above are integrated using softmax voting.

V. Experimental Setup and Results

The proposed ensemble of deep learning models-based violin bowing technique classifier was trained and tested on 4 datasets *primary dataset*, *dynamic dataset* and *tempo dataset* and *full dataset* using 10-fold stratified cross-validation, 9-folds for training and 1-fold for testing.

A. Primary Dataset

The results of applying ensemble of deep learning models classifier for the *primary dataset* is shown in Table III.

TABLE III. Classification Metric of Ensemble of Deep Learning Models for Primary Dataset

	Precision	Recall	F1-Score	Support
detaché	1.00	0.99	0.99	80
legato	0.96	1.00	0.98	80
martelé	0.99	0.99	0.99	80
spiccato	0.99	1.00	0.99	80
staccato	1.00	0.96	0.98	80
macro avg	0.99	0.99	0.99	400
weighted avg	0.99	0.99	0.99	400

Fig. 9 shows the confusion matrices of CNN audio classifier, 2D LSTM-EMG motion classifier, 2D LSTM-IMU motion classifier and ensemble model classifier for *primary dataset*.

For *primary dataset*, the accuracies of applying CNN using audio features, 2D LSTM using EMG-motion features and 2D LSTM using IMU-motion features are 97.5%, 90.5% and 95.0%, respectively. As it can be seen in the confusion matrices Fig. 9, "legato" is classified better with the audio classifier and "martelé" is classified better with the IMU-motion classifier. With the proposed ensemble of deep learning models, the accuracy obtained is 98.8%.

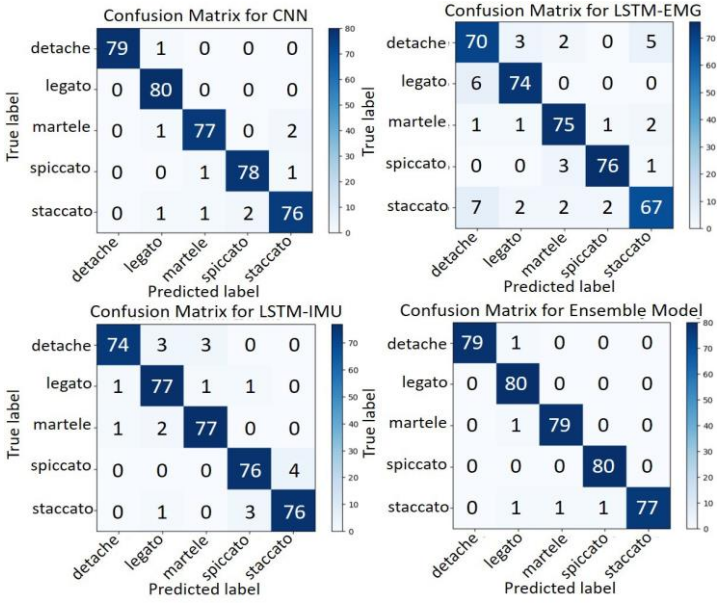


Fig. 9. Confusion matrices of primary dataset.

B. Dynamic Dataset

The results of applying ensemble of deep learning model classifier for the *dynamic dataset* is as shown in Table IV. Fig. 10 shows the confusion matrices of CNN audio classifier, 2D LSTM-EMG motion classifier, 2D LSTM-IMU classifier and ensemble model classifier for *dynamic dataset*.

For *dynamic dataset*, the accuracies of applying CNN using audio features, 2D LSTM using EMG-motion features and 2D LSTM using IMU-motion features are 97.5%, 91.7%, and 99.2%, respectively. With the proposed ensemble of deep learning models, the accuracy obtained ~100.0%. This indicates that the ensemble of deep learning models is able to classify and predict bowing techniques of violin excerpts with variation in dynamics, i.e., from soft to loud, with greater accuracy.

TABLE IV. CLASSIFICATION METRIC OF ENSEMBLE OF DEEP LEARNING MODEL FOR DYNAMIC DATASET

	Precision	Recall	F1-Score	Support
detaché	1.00	1.00	1.00	80
legato	1.00	1.00	1.00	80
spiccato	1.00	1.00	1.00	80
macro avg	1.00	1.00	1.00	240
weighted avg	1.00	1.00	1.00	240

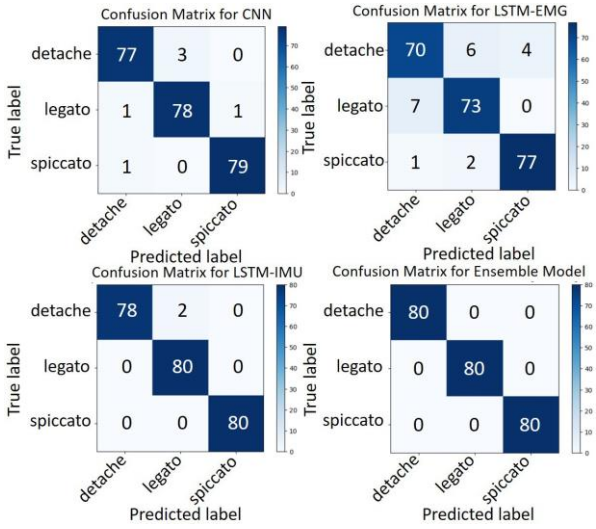


Fig. 10. Confusion matrices of dynamic dataset.

C. Tempo Dataset

The results of applying ensemble of deep learning model classifier for the *tempo dataset* is shown in Table V. Fig. 11 shows the confusion matrix of CNN audio classifier, 2D LSTM-EMG motion classifier, 2D LSTM-IMU classifier and ensemble model classifier for *dynamic dataset*.

TABLE V. CLASSIFICATION METRIC OF ENSEMBLE DEEP LEARNING MODEL FOR TEMPO DATASET

	Precision	Recall	F1-Score	Support
detaché	1.00	0.99	0.99	80
legato	0.99	0.99	0.99	80
spiccato	0.99	1.00	0.99	80
macro avg	0.99	0.99	0.99	240
weighted avg	0.99	0.99	0.99	240

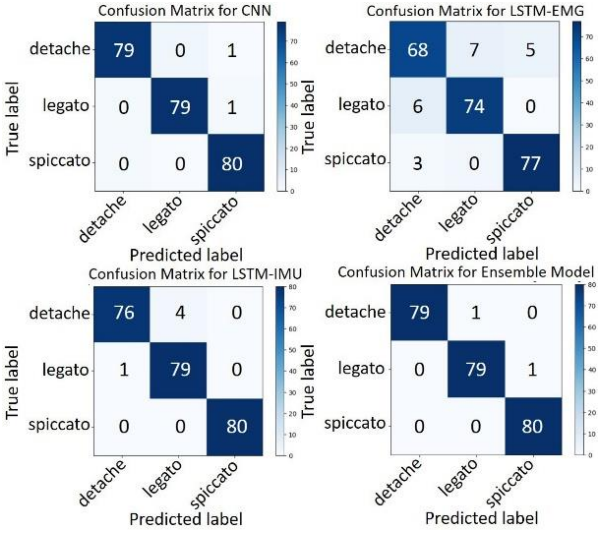


Fig. 11. Confusion matrices of tempo dataset.

For *tempo dataset*, the accuracies of applying CNN using audio features, 2D LSTM using EMG-motion features and 2D LSTM using IMU-motion features are 99.2%, 91.2%, and 97.9%, respectively. With the proposed ensemble of deep learning model, the accuracy obtained is 99.2%. This indicates that the ensemble of deep learning models is able to classify and predict bowing techniques of violin excerpts with variation in tempo, i.e., from slow to fast speed, with greater accuracy.

D. Full Dataset

The results of applying ensemble of deep learning model classifier for the *full dataset* is as shown in Table VI. Fig. 12 shows the confusion matrices of CNN audio classifier, 2D LSTM-EMG motion classifier, 2D LSTM-IMU motion classifier and ensemble model classifier for *dynamic dataset*. For *full dataset*, the accuracies of applying CNN using audio features, 2D LSTM using EMG-motion features and 2D LSTM using IMU-motion features are 99.0%, 89.5%, and 96.7%, respectively. With the proposed ensemble of deep learning models, the accuracy obtained is 99.5%. This shows that the ensemble of deep learning models is able to classify and predict the samples of violin excerpts with variation in dynamics and variation in tempo to the respective classes without much uncertainty. That is, violin excerpts with the "detaché" technique with soft-loud variation are classified as "detaché" class. Similarly, violin excerpts with the "detaché" technique with slow-fast variation are classified as "detaché" class.

TABLE VI. CLASSIFICATION METRIC OF ENSEMBLE OF DEEP LEARNING MODELS FOR FULL DATASET

	Precision	Recall	F1-Score	Support
detaché	1.00	0.99	1.00	240
legato	0.99	1.00	1.00	240
martelé	0.99	0.99	0.99	80
spiccato	1.00	1.00	1.00	240
staccato	0.99	1.00	0.99	80
macro avg	0.99	1.00	0.99	880
weighted avg	1.00	1.00	1.00	880

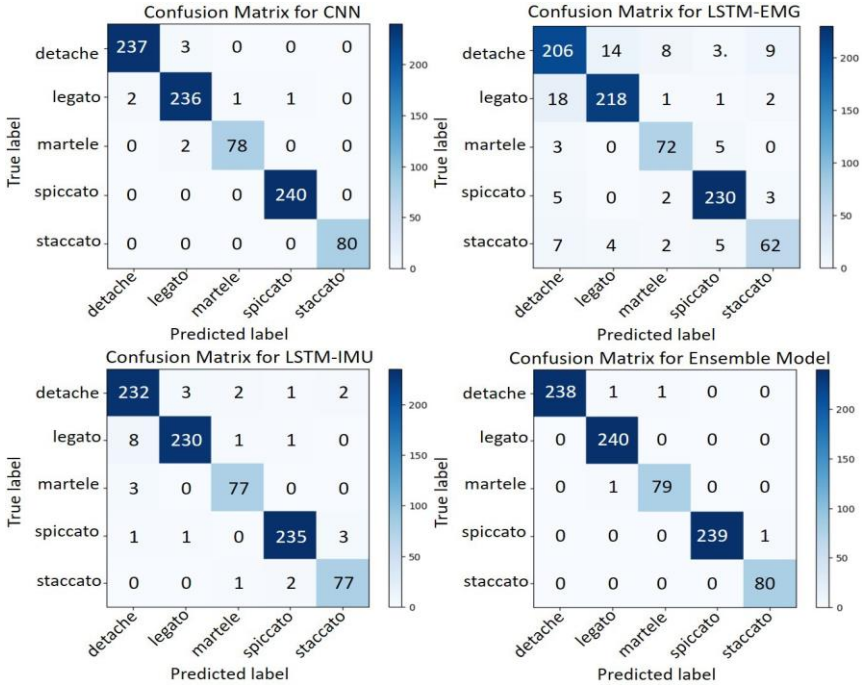


Fig. 12. Confusion matrices of full dataset.

TABLE VII. COMPARISON OF ALVARO *ET AL*. [3] AND PROPOSED METHOD

Datasets	Alvaro *et al.* [3]				Proposed Method			
	PF-IMU	HMM5-IMU	PF-IMU+EMG	HMM5-IMU+EMG	1D CNN-AUDIO	2D LSTM-EMG	2D LSTM-IMU	Softmax Voting
Primary	91.3	98.9	94.3	90.2	97.5	90.5	95	98.8
Dynamic	81.6	81.3	41.2	39.5	97.5	91.7	99.2	100
Tempo	82	81.2	40.7	37	99.2	91.2	97.9	99.2
Full	-	-	-	-	99	89.5	96.7	99.5

The comparison of accuracies of the proposed approach with the previous study Alvaro *et al*. [3] is shown in Table VII. In the proposed approach the accuracies obtained are consistent among all four datasets. It can be seen that in the previous study [3] the accuracies of *dynamic* and *tempo datasets* have decreased compared to the *primary dataset* and when both IMU and EMG are considered, the accuracies are very less.

TABLE VIII. COMPARISON OF PROPOSED METHOD WITH EXISTING WORKS

Features	[3]	[10]	[11]	[12]	[13]	[14]	Proposed Method
Audio	-	94.8	93.2	-	-	-	99.0
Motion	94.5	-	39.1	31.3	98.31	95.0	96.7
Audio & Motion	-	-	94.6	-	-	-	99.5
Camera Depth	-	-	-	74.7	-	-	-
Camera Depth & Motion	-	-	-	80.9	-	-	-

The comparisons of accuracies of the proposed approach with the previous studies [3, 10–14] are shown in Table VIII. It can be observed that the overall accuracies of the proposed approach considering either audio, motion or audio and motion features are better than previous studies. Since Dalmazzo *et al*. [13] has considered a simple G-Major scale violin excerpt and the accuracy may be slightly higher for the motion-only classifier.

VI. CONCLUSION

For identifying the bowing technique both the sound produced by the violin and the movement on the right arm is essential. In this paper, an investigation of violin bowing technique classification using audio features from Essentia [15] and time domain motion features was done and evaluated on four datasets namely *primary*, *dynamic*, *tempo* and *full datasets*. Better classification of "legato" in the audio-classifier and "martelé" in the motion-classifier has driven the proposal of ensemble of deep learning, 1-CNN for audio features and 2-LSTM for EMG-motion and IMU-motion features, for violin bowing technique classification. The results show that the use of both audio and motion features gives better accuracy.

Violin bowing technique classification accuracy of 98.8%, 100.0%, 99.2%, 99.5% obtained for *primary*, *dynamic*, *tempo* and *full datasets* respectively with ensemble of deep learning models is better than the earlier proposed approaches to the best of our knowledge.

In future, we would like to study the extraction of frequency domain features from motion signals and use them to train and predict violin bowing techniques, that may further give better accuracy. Furthermore, we would

like to study and improve the model that can classify more violin bowing techniques.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHORS CONTRIBUTION

All authors participated in the research, experiments and analysis of data. Zain Muhammed, Nagamanoj Karunakaran, and Pranamya P. Bhat carried out the implementation. Arti Arya helped with the logic of code. All authors discussed the results and contributed equally on manuscript and finally all authors proof read on individual basis and collectively. All authors had approved the final version.

REFERENCES

[1] Violin Masterclass: The Sassmannshaus tradition for violin playing. [Online]. Available: http://www.violinmasterclass.com/

[2] N. Karunakaran and A. Arya, "A scalable hybrid classifier for music genre classification using machine learning concepts and spark," in *Proc. 2018 International Conference on Intelligent Autonomous Systems (ICoIAS)*, 2018, pp. 128–135. doi: 10.1109/ICoIAS.2018.8494161

[3] A. Sarasua, B. Caramiaux, A. Tanaka, and M. Ortiz, "Datasets for the analysis of expressive musical gestures," in *Proc. 4th International Conference on Movement Computing*, Jun 2017, London, United Kingdom, pp. 1–4.

[4] C. Tepe and M. C. Demir, "Detection and classification of muscle activation in EMG data acquired by Myo armband," *European Journal of Science and Technology*, pp. 178–183, 2020.

[5] C. Tepe and M. C. Demir, "Real-time classification of EMG Myo armband data using support vector machine," *Innovation and Research in BioMedical Engineering*, vol. 43, issue 4, pp. 300–308, 2022. https://doi.org/10.1016/j.irbm.2022.06.001

[6] M. Georgi, C. Amma, and T. Schultz, "Recognizing hand and finger gestures with IMU based motion and EMG based muscle activity sensing," in *Proc. the International Conference on Bio-inspired Systems and Signal Processing*, 2015, pp. 99–108. doi: 10.5220/0005276900990108

[7] H. Ali and W. Yanen, "SVM classification for novel time domain IMU and EMG fused features for control of 6-DOF industrial robot," in *Proc. 2020 IEEE International Conference on Mechatronics and Automation (ICMA)*, 2020, pp. 18–22. doi: 10.1109/ICMA49215.2020.9233536.

[8] D. Xiong, D. Zhang, X. Zhao, and Y. Zhao, "Deep learning for EMG-based human-machine interaction: A review," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 3, pp. 512–533, March 2021. doi: 10.1109/JAS.2021.1003865

[9] J. P. Vásconez, L. I. B. López, Á. L. V. Caraguay, and M. E. Benalcázar, "Hand gesture recognition using EMG-IMU signals and deep Q-networks," *Sensors*, vol. 22, no. 24, 9613, 2022. doi: 10.3390/s22249613

[10] D. Mukherjee and D. Arvind, "The speckled cellist: Classification of cello bowing techniques using the orient specks," *EAI Endorsed Trans. Perv. Health Tech.*, vol. 2, no. 6, 3, Dec. 2015.

[11] D. Dalmazzo and R. Ramırez, "Bowing gestures classification in violin performance: A machine learning approach," *Frontiers in Psychology*, vol. 10, 344, 2019.

[12] S.-W. Sun, B.-Y. Liu, and P.-C. Chang, "Deep learning-based violin bowing action," *Sensors*, vol. 20, no. 20, 5732, 2020. https://doi.org/10.3390/s20205732

[13] D. Dalmazzo, G. Waddell, and R. Ramırez, "Applying deep learning techniques to estimate patterns of musical gesture," *Front. Psychol.*, vol. 11, 575971, 2021. doi: 10.3389/fpsyg.2020.575971

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

[15] D. Bogdanov, N. Wack, E. Gomez, S. Gulati, P. Herrera1, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, "ESSENTIA: An audio analysis library for music information retrieval," in *Proc. 21st ACM International Conference on Multimedia*, 2013, pp. 855–858

[16] J. S. Hussain, A. Al-Khazzar, and M. N. Raheema, "Recognition of new gestures using Myo armband for myoelectric prosthetic applications," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 6, pp. 5694–5702, December 2020. doi: 10.11591/ijece.v10i6.pp5694-5702

[17] H. A. Javaid, M. I. Tiwana, A. Alsanad, J. Iqbal, M. T. Riaz, S. Ahmad, and F. A. Almisned, "Classification of hand movements using Myo armband on an embedded platform," *Electronics*, vol. 10, 1322, 2021. https://doi.org/10.3390/ electronics10111322

[18] M. Arozi, M. Ariyanto, A. Kristianto, Munadi, and J. D. Setiawan, "EMG signal processing of Myo armband sensor for prosthetic hand input using RMS and ANFIS," in *Proc. 2020 7th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, 2020, pp. 36–40. doi: 10.1109/ICITACEE50144.2020.9239169

[19] J. Lopes, M. Simao, N. Mendes, M. Safeea, J. Afonso, and P. Neto, "Hand/arm gesture segmentation by motion using IMU and EMG sensing," in *Proc. 27th International Conference on Flexible Automation and Intelligent Manufacturing*, Modena, Italy, 2017.

[20] M. F. Rabbi, K. H. Ghazali, N. U. Ahamed, and T. Sikandar, "Time and frequency domain features of EMG signal during Islamic prayer (Salat)," in *Proc. 2017 IEEE 13th International Colloquium on Signal Processing Its Applications (CSPA)*, 2017, pp. 139–143. doi: 10.1109/CSPA.2017.8064939