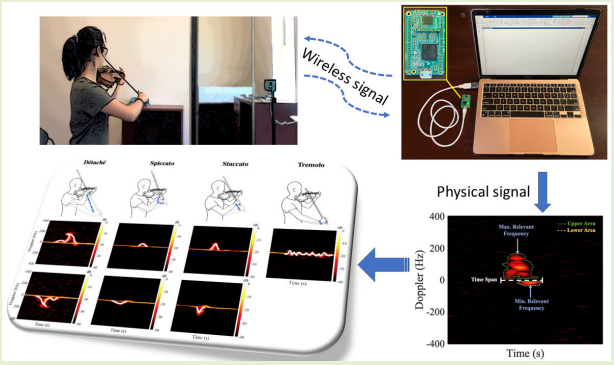
Automated Violin Bowing Gesture Recognition Using FMCW-Radar and Machine Learning

Hannah Gao and Changzhi Li, *Senior Member, IEEE*

***Abstract*—A key step in learning the violin is mastering control over various bowing techniques since the drawing of the violin bow directly influences the sound quality produced. As it is important for violinists to receive frequent feedback on their bowing motions, there is a need for digital means of providing automated feedback to musicians. This study uses a 60-GHz frequency-modulated-continuous-wave (FMCW) radar to gather data on the violinist's bowing arm for a total of seven bowing gestures: détaché up, détaché down, spiccato up, spiccato down, staccato up, staccato down, and tremolo. A total of 1200 bowing gestures for three different violinists are recorded using radar. The raw signal data from the radar is processed to generate time-Doppler spectrograms of the gestures. Features are extracted from the time-Doppler data using two different methods and fed into machine-learning models for the automated classification of bowing gestures. The first method involves manually engineering features to be extracted from the signal data matrix. The second method leverages the power of convolutional neural networks (CNNs) to automatically extract features from images of the time-Doppler spectrograms. Comparing model performances reveals that fine-tuning a pretrained SqueezeNet CNN model yields the highest classification accuracy (95.00%). This study also analyzes the influence of fluctuations in the overall user-to-radar range on the time-Doppler spectrograms produced.**



***Index Terms*—Bowing gestures, feature extraction, frequency-modulated-continuous-wave (FMCW) radar, machine learning, time-Doppler, violin.**

I. Introduction

LEARNING the violin is a complicated task requiring thousands of hours of practice, especially given the difficulty of drawing out a pleasant sound from the instrument [1]. Since the violin bow's interaction with the strings directly influences the sound produced, gaining control of different bowing motions is the first key step in learning the instrument. By altering the amount of pressure exerted on the bow, the bow speed, and the positioning of the bow on the strings, one can directly alter the quality of the tone produced [2]. Thus, receiving feedback on bowing techniques is a priority for beginner to intermediate violin students. Although music teachers can provide such input, private instructors can be costly and may be unaffordable for individuals to pursue long-term. Furthermore, even with an instructor, most of the hours spent developing the necessary sensorimotor skills occur outside of the lesson time [3]. During the time gaps, feedback is necessary to help beginning students reinforce the general bowing strokes introduced during lessons and to identify the consistency with which their stroke matches that marked in the music. Because of the necessity for affordable and continuous performance feedback, technology is being increasingly exploited for the automated analysis of bowing techniques.

Beyond the purpose of providing direct feedback to the user, however, technology-aided analysis of violin bowing techniques has other pedagogical applications such as enhancing interactive music training interfaces. For instance, technology-based recognition of bowing gestures in real-time is necessary for music-training software to automatically mark up scores while the user is performing or to track the violinist's location in the music when using play-along tools [4].

The task of detecting and/or recognizing bowing motions can be approached in several different ways. For example, a popular method in previous studies is attaching electronics to different parts of the instrument to gather information about

Manuscript received 22 March 2023; accepted 25 March 2023. Date of publication 5 April 2023; date of current version 1 May 2023. This work was supported by the National Science Foundation (NSF) under Grant ECCS-2030094 and Grant ECCS-1808613. The associate editor coordinating the review of this article and approving it for publication was Dr. Fabrizio Santi. *(Corresponding author: Changzhi Li.)*

Hannah Gao is with the Harriton High School, Rosemont, PA 19010 USA (e-mail: hngao2014@gmail.com).

Changzhi Li is with the Department of Electrical and Computer Engineering, Texas Tech University, Lubbock, TX 79409 USA (e-mail: changzhi.li@ttu.edu).

Digital Object Identifier 10.1109/JSEN.2023.3263513

the bow's acceleration, position, speed, strain, and so on. Such information can then be used to differentiate bowing techniques and evaluate their accuracy. Sensors such as strain gauges, accelerometers, and electromagnetic field sensors are used in new violin interfaces like the Hyperbow [5], [6] and other "augmented instruments" [7]. Some studies make use of vision systems for automated violin bowing feedback, employing depth cameras and video streams to capture general limb motions of the user, which is often used in conjunction with sensors to capture more subtle movements [8], [9], [10]. Others analyze audio signals to identify various bowing characteristics and predict the bowing technique being performed [11], [12]. Several studies use wearable devices to track the motion of the violinist's body parts rather than the instrument itself. Dalmazzo and Ramírez [13] equip violinists with bracelets around the forearm to track electrical muscle activity while Linden et al. [14] position sensors on the users' arms and torso.

However, each of these approaches to automating musical instrument instruction has limitations. Sensor attachments and wearable devices can feel uncomfortable and invasive for users. These electronics also add considerable bulk to the instrument, making it difficult for the musician to maneuver the violin bow normally and with flexibility [15]. Vision systems may infringe on privacy and are sensitive to lighting conditions as well as clothing that conceals precise limb movements. Finally, audio signal analysis is sensitive to the unique properties of different violins and bows as well as noise in the sound frequency; furthermore, audio signals do not directly provide any information on the hand and arm movements of the user.

A radar system is a non-contact method of sensing that preserves user privacy and can sense through clothing to capture subtle limb movements. Radars show great promise for motion detection in a wide range of applications. Oyamada et al. [16], Wang et al. [17], and Choi et al. [18] used radar systems to track body displacement for vital signs monitoring. Li et al. [19] analyzed radar micro-Doppler and range-Doppler information for differentiating concealed rifle carrying from other similar activities that may trigger a false alarm. Radars also show the potential in recognizing different hand gestures for human–computer interaction [20], [21].

In this study, a portable frequency-modulated continuous-wave (FMCW) radar is used to detect different violin bowing gestures. FMCW radars emit a continuous signal whose frequency is gradually increased during each period of measurement. By comparing the transmitted and received signals, information about the target's range and velocity can be determined. In a preliminary work, presented at the IEEE Topical Conference on Wireless Sensors and Sensor Networks [22], the time-Doppler spectrograms and range profiles for four standard violin bowing techniques are visually analyzed and compared to incorrect bowing techniques. This study builds on the previous work by providing further analysis of the time-Doppler spectrograms of the four bowing techniques and employing machine learning for the automated recognition of seven bowing gestures. Moreover, the impact of user-to-radar range fluctuations on the reliability of radar-based violin pedagogy is analyzed.

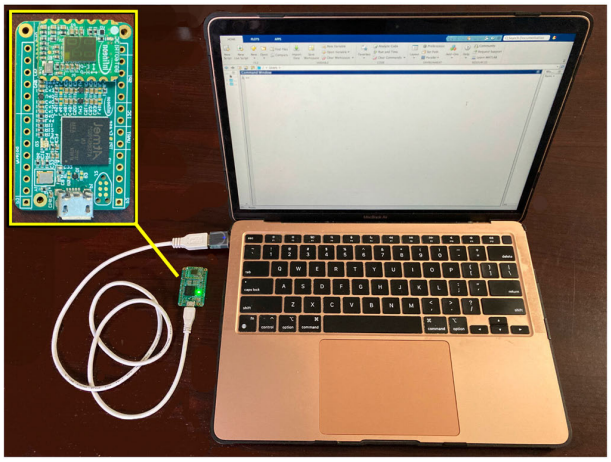


Fig. 1. Complete sensing system based on a 60-GHz Infineon FMCW radar module and a laptop.

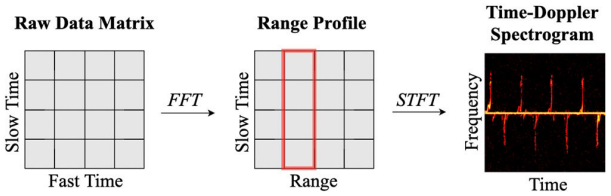


Fig. 2. Flowchart showing signal processing steps from the raw data matrix to the time-Doppler spectrogram.

II. Theory of Radar-Based Gesture Recognition

Fig. 1 shows the sensing system employed in this study, which is based on a 60-GHz Infineon radar module and a laptop. The raw signal received by the radar's receiver is processed to generate time-Doppler spectrograms for the following bowing techniques: détache, spiccato, staccato, and tremolo. The bowing techniques are described in Table I. Each of the bowing techniques except for tremolo can be further classified into an up-bow gesture (when the hand draws the bow in an upward motion) and a down-bow gesture (when the hand draws the bow in a downward motion), for a total of seven bowing gestures to be classified. The tremolo bow stroke rapidly oscillates between up and down motions and should not be further classified into up-bow or down-bow gestures. Time-Doppler information for each gesture is fed into machine-learning models for automated classification.

A. Signal Processing

To collect data on the user's performance, the FMCW radar transmits a frequency-modulated signal, or chirp, which reflects off the violinist's arm and hand and is received by the radar's receiving antenna for processing. The signal processing flowchart is shown in Fig. 2. The beat signal detected by the mixer of the radar's receiver chain is stored as voltage readings in a matrix. Each row in the matrix represents a single period of measurement that corresponds to a transmitted linear frequency-modulated chirp. A time-range map can be generated by performing a fast Fourier transform (FFT) along the fast time of the data matrix, represented as [23]

$$S_b(f) = \sigma T \exp\left(j\frac{4\pi f_c R(\tau)}{c}\right) \operatorname{sinc}\left(T\left(f - \frac{2\gamma R(\tau)}{c}\right)\right) \quad (1)$$

TABLE I
BOWING TECHNIQUE DESCRIPTIONS

Bowing Technique	Description
Détaché	Separate bow strokes played smoothly and without accent
Spiccato	Light, bouncing bow strokes
Staccato	Short, detached bow strokes played on the strings
Tremolo	Tiny and rapid up and down bow movement, typically performed at the bow tip

Brief bowing technique descriptions from [19].

where $S_b(f)$ is the frequency domain representation of the beat signal, σ is the baseband signal amplitude, T is the chirp duration, f_c is the center frequency, τ is slow time, $R(\tau)$ is the variation in the target's range relative to the radar board, and c is the speed of light. To extract the time-Doppler spectrogram, the range-of-interest is isolated in the time-range map, and a short-time Fourier transform (STFT) is applied along the slow time.

B. Features Extracted From the Time-Doppler Spectrograms

Selecting features to extract from the time-Doppler signature is an important step in training a machine-learning model to accurately predict the bowing gesture being performed. In this work, two different methods of feature extraction are explored: 1) manual feature extraction and 2) automated feature extraction using a convolutional neural network (CNN). One of the main benefits of CNN models is their ability to automatically discern important features in the given data without any human assistance. Although using CNNs to automatically extract features is a very popular approach to image classification tasks, manually selecting and extracting features for more simple machine learning models can, in some cases, yield similar or better results [24]. This section will describe the manually selected features.

Features of the time-Doppler spectrograms that vary significantly between at least two types of gestures are selected, as they are most useful for differentiating between the spectrograms of various gesture classes. The selected features are extracted from the power matrices representing the time-Doppler spectrograms. Each row of the matrix stores power values (dB) for a particular Doppler frequency over time. Positive Doppler frequency values correspond to movement of the target away from the radar, typical of up-bow gestures where the user draws the bow in toward the violin body; conversely, negative Doppler frequency values correspond to movement of the target toward the radar. Prior to extracting the features, the baseband region (from -13 to $+13$ Hz) and anything with normalized power below -35 dB is removed (i.e., set to -1000 dB). Entries with low power values can be considered noise as they are unlikely to be a meaningful part of the signature for the bowing gesture being performed. In this work, the following features are extracted from each 2-s window of time-Doppler signatures for a single bow stroke (see Fig. 3).

1) *Maximum Relevant Frequency (F_max):* Call a particular row of the normalized power matrix *non-empty* if its maximum entry is not equal to -1000 dB (i.e., every entry is at least -35 dB). *F_max* is found by starting at $+13$ Hz and iterating through successively higher frequencies, stopping when a frequency, f, meeting the following criteria is reached: 1) the row of the matrix corresponding to f is nonempty and 2) the difference between f and last frequency with a nonempty row is no more than 25 Hz. A margin of 25 Hz accounts for possible gaps in the gesture's signature created when removing all entries with power under -35 dB.
2) *Minimum Relevant Frequency (F_min):* The algorithm for this feature is the same as that of *F_max* except for the iteration that starts at a frequency of -13 Hz and moves through successively lower frequencies.
3) *Upper Area (A_upper):* This feature is calculated by traversing through all rows representing frequencies greater than $+13$ Hz and counting the number of entries above -35 dB.
4) *Lower Area (A_lower)*: The algorithm for this feature is the same as that of *A_upper* except all rows representing frequencies lower than -13 Hz are searched.
5) *Upper Power (P_upper):* This feature is calculated by traversing through all rows representing frequencies greater than $+13$ Hz and summing all entries above -35 dB.
6) *Lower Power (P_lower):* The algorithm for this feature is the same as that of *P_upper* except all rows representing frequencies below -13 Hz are searched.
7) *Time Span (T_span):* The feature represents the total amount of time the signature spans. Call a particular column, or time, of the power matrix *non-empty* if its maximum entry is not equal to -1000 dB. The time span is calculated by counting the number of non-empty columns.

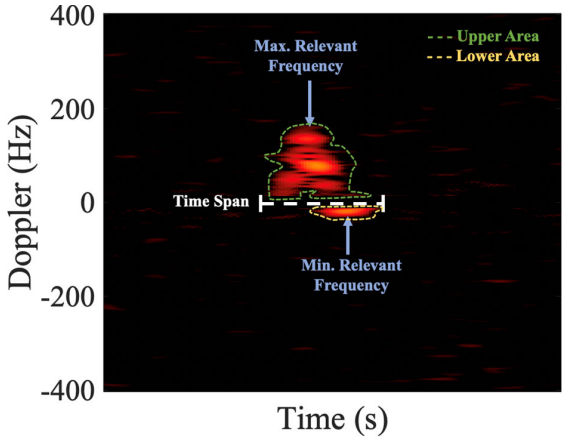


Fig. 3. Illustration of the maximum relevant frequency, minimum relevant frequency, upper area, lower area, and time span features for staccato up-bow.

III. EXPERIMENTAL SETUP AND DATA COLLECTION

This study uses an Infineon 60-GHz FMCW radar with a center frequency of 60.5 GHz, a bandwidth of 5 GHz, and a sampling rate of 2 MHz to collect data on bowing motions. The radar is used to collect data on the seven bowing gestures from three different subjects. See Table II for basic subject

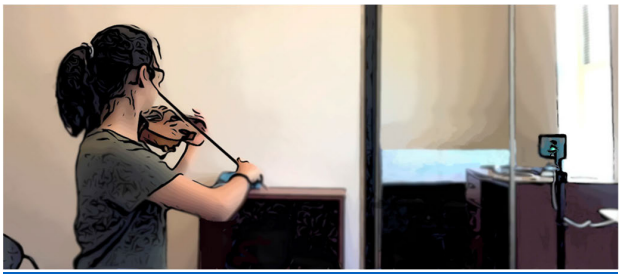


Fig. 4. Experimental setup.

TABLE II
SUBJECT DESCRIPTIONS

Subject	Gender	Height (cm)	Arm Length (cm) [a]
1	F	161	64
2	M	180	74
3	M	173	75

[a] Arm length is defined as the distance from the shoulder to the furthest fingertip.

descriptions. Each subject is seated on a chair 1 m from the radar to minimize the amount of lower body movement picked up by the radar. The violin is held parallel to the radar board and the radar antenna's height is adjusted to be level with the violin. The configuration is shown in Fig. 4.

A total of 1200 time-Doppler data samples are collected to train machine learning classifiers. Subjects are asked to perform ten sets of each bowing technique on the A string alongside a 60 beats/min metronome, with each set lasting for 20 s. For détaché, spiccato, and staccato, each set consists of five up-bows and five down-bows performed consecutively, with each stroke falling on every other click of the metronome. For all subjects combined, this yields a total of 150 data samples for each of the following bowing gesture classes: détaché up-bow, détaché down-bow, spiccato up-bow, spiccato down-bow, staccato up-bow, and staccato down-bow. For the tremolo bowing technique, there is no distinct up- or down-bow stroke; each set simply consists of ten tremolo oscillations. Among the three subjects, this yields a total of 300 data samples for the tremolo gesture class.

IV. VIOLIN BOWING GESTURE CLASSIFICATION

Two different approaches are used to classify the time-Doppler spectrograms using machine learning. The first approach trains models on the radar signal data, using the manually extracted features described in Section II. The second approach trains models on images of the time-Doppler spectrograms, using the ability of a CNN to automatically extract features. This second approach can be further divided into two different sub-approaches: 1) performing both feature extraction and fine-tuning on a pretrained CNN and 2) using the initial feature-learning layers of the pretrained CNN as the backbone and feeding the automatically learned features into a separate non-CNN classifier.

The following types of machine learning classifiers are used in this study: support vector machine (SVM), decision tree, k-nearest neighbor (KNN), linear discriminant analysis (LDA), multilayer perceptron (MLP), and (CNN). For all these models, the bowing gesture data is split into training and testing sets in a 70:30 ratio.

Support Vector Machine (SVM): This algorithm seeks the hyperplane, or boundary, that fully separates the data points into each of their classes and maximizes the distance between the hyperplane and support vectors (the data points situated closest to the hyperplane) [25]. This work uses the popular Gaussian kernel function.

Decision Tree: This algorithm can be represented by a tree structure where features are represented by nodes that branch out into sub-nodes representing decisions [26]. Decision trees classify inputs by starting at the root node and progressing lower in the tree by taking decision paths down the decision nodes until a leaf node is reached.

K-Nearest Neighbor (KNN): This algorithm assumes the proximity of data points with similar features. For each observation fed to the classifier, the class with the greatest representation among the K-nearest data points is the output label [27].

Linear Discriminant Analysis (LDA): This algorithm projects data into a lower dimension with the goal of maximizing the distance between class means while minimizing variance within classes. To make predictions, LDA uses Bayes' Theorem to calculate the probability of a sample belonging to each of the classes and outputs the highest probability class.

Multilayer Perceptron (MLP): MLP is a feedforward artificial neural network (ANN) composed of fully connected layers [28].

Convolutional Neural Network (CNN): A CNN is a specific type of neural network that contains at least one convolutional layer [29]. These models are particularly useful for classifying 2-D image inputs and can automatically extract key features without human assistance.

A. Using Manually Extracted Features

A popular method of classifying gestures from radar data is training machine learning classifiers on features extracted from the time-Doppler data matrix [30]. In this study, the manually extracted features from the power data matrix are fed into the following machine learning models: SVM, decision tree, KNN, LDA, and MLP. The MLP has two fully connected layers and uses a rectified linear unit (ReLU) activation function for the first fully-connected layer and a SoftMax function for the final layer.

B. Using Automatically Extracted Features

While manual feature extraction can provide greater interpretability of model results, it requires a solid understanding of the domain and the data to best capture patterns. Automatic feature extraction, meanwhile, is an easier and more efficient way to pick up patterns in the data [31]. This study exploits the automatic feature extraction ability of CNNs to classify images of the time-Doppler spectrograms of bowing gestures in two ways: 1) using transfer learning and 2) using a CNN to extract features for training another machine learning model. Both methods make use of already-learned knowledge contained in a pretrained CNN, making the process faster and less

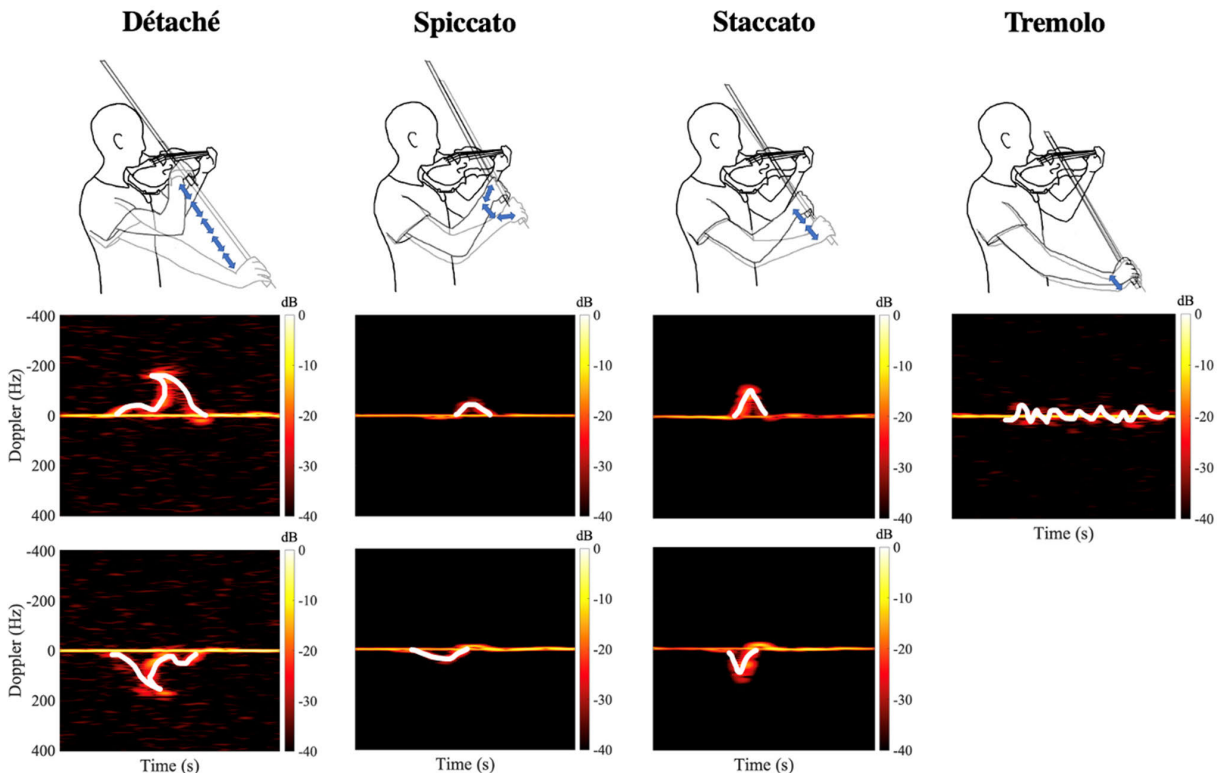


Fig. 5. Sketches of bowing techniques and their corresponding up-bow and down-bow time-Doppler spectrograms. Bowing technique sketches are reused from [19].

computationally heavy than training a CNN from scratch. This work uses the learned features from a SqueezeNet CNN model from MATLAB's Deep Learning Toolbox which was pretrained on image data from *ImageNet* [32].

1) *Transfer Learning:* Transfer learning involves adjusting a pretrained network for a new task by replacing the final layers of the model and re-training the model to update the weights. By setting the learning rate in the new final convolutional layer to be greater than that in the transferred layers, the model makes minimal changes to the previously learned weights and instead performs most of the updates in the final layers. This allows for efficient training of a CNN with a relatively small dataset for the new task. The fine-tuning of the SqueezeNet is performed in MATLAB using a learning rate of 0.001, a mini-batch size of 64, and a total of 16 epochs. A train-validation-test split ratio of 70:20:10 is used in this approach. In other words, 70%, 20%, and 10% of the samples from each class are selected randomly to serve as the training, validation, and testing sets, respectively, without regard to the subject performing the gesture.
2) *Backbone CNN for Feature Extraction:* The pretrained SqueezeNet model has already learned important image features from the large *ImageNet* dataset and can serve as a backbone for training non-CNN models. Features, such as edges, learned in the earlier layers of the backbone model are generally applicable to images in a variety of other tasks; conversely, features learned in layers closer to the output layer of the backbone model are more specific to the task the model was trained on. Features extracted from the backbone CNN model are applied to the time-Doppler images, and these features are fed into the following machine-learning models for classification: SVM, decision tree, KNN, LDA, and MLP.

V. Results

Fig. 5 shows 2-s time-Doppler spectrograms for down-bow (top spectrogram) and up-bow (bottom spectrogram) motions for the détaché, spiccato, staccato, and tremolo bowing techniques. The sketches highlight the most energetic part of typical spectrograms measured from these bowing techniques. It can be observed that, in ideal cases, the signatures for all the bowing techniques are easily distinguishable from one another by analyzing the features described in Section II. Down-bow and up-bow gestures can be distinguished through the *A_upper* and *A_lower* features: if *A_upper* is significantly greater than *A_lower*, an up-bow is being performed; if *A_upper* is significantly less than *A_lower*, a down-bow is being performed; if *A_upper* and *A_lower* are roughly equal, then a tremolo, which has no distinct up- or down-bow, is being performed. Similarly, *P_upper* and *P_lower* can be used to help differentiate up-bow and down-bow gestures. *F_max* and *F_min* for spiccato and tremolo are smaller in magnitude compared to détaché and staccato. *T_span* is the greatest for tremolo and détaché gestures as they span longer periods of time than spiccato and staccato.

The bowing gesture prediction accuracy for all machine learning classifiers is displayed in Table III. The accuracy is calculated as the total number of correct predictions across all classes divided by the total size of the dataset. Among all the classifiers trained on the manually extracted features, the

TABLE III
MACHINE LEARNING RESULTS

Feature-Extraction Method	Machine Learning Classifier	Accuracy
Manual	SVM	84.72%
	Decision Tree	79.44%
	KNN	77.22%
	LDA	75.56%
	MLP	85.28%
Automated	SVM	85.28%
	Decision Tree	50.83%
	KNN	70.83%
	LDA	89.44%
	MLP	82.78%
	SqueezeNet CNN	95.00%

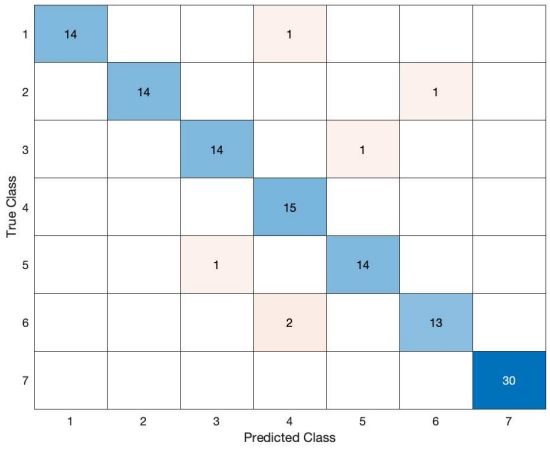


Fig. 6. Confusion matrix for the SqueezeNet CNN model. Numbers 1–7 represent the classes in the following order: détaché down, détaché up, spiccato down, spiccato up, staccato down, staccato up, and tremolo.

MLP model achieves the highest accuracy of 85.28%, closely followed by the SVM. Among the classifiers trained on features automatically extracted by a SqueezeNet CNN, the best performance (95.00% accuracy) is achieved by fine-tuning the CNN model itself. Overall, the fine-tuned SqueezeNet CNN achieves the highest accuracy across all models; however, it is important to recognize the substantial increase in training time required of this approach.

Fig. 6 displays the confusion matrix for the fine-tuned SqueezeNet CNN model. The détaché up-bow, spiccato up-bow, staccato up-bow, and tremolo gestures are consistently classified correctly. The staccato up-bow gesture is the most frequently misclassified.

Fig. 7 displays the prediction accuracies within each gesture type for all machine learning classifiers. In general, models are most successful in classifying tremolo and detaché up-/down-bows. Models tend to struggle the most with classifying spiccato and staccato up- and down-bows, likely due to the very similar Doppler spectrograms produced by the two gestures. The SqueezeNet CNN model outperforms all other models in classifying each type of gesture except for the detaché up-bow and staccato up-bow strokes. However, the CNN model is not outperformed significantly in these classes. The MLP (A) model's accuracy exceeds that of the CNN by 3% when classifying détache up-bows. The decision tree (M) and MLP (M) models' accuracies only exceed that of the CNN by 2% when classifying staccato up-bows.

TABLE IV
MODEL INFERENCE TIMES

Classification Method	Average Single-Sample Inference Time (x 10^{-6} s)
SVM (M)	111.9
Decision Tree (M)	10.7
KNN (M)	18.1
LDA (M)	16.9
MLP (M)	4.0
SVM (A)	190.1
Decision Tree (A)	14.2
KNN (A)	488.4
LDA (A)	118.8
MLP (A)	15.3
SqueezeNet CNN (A)	25800.0

Although the fine-tuned CNN performs better than the other models, it is important to consider the large increase in computation time that comes with using complex neural networks. To gauge the difference in computation times of various classification methods, the average single-sample inference time is calculated for each model by averaging the inference times across 100 passes through the testing dataset; within a single pass, the inference time is taken as the mean of the single-sample inference times across all samples. Table IV displays the average single-sample inference times for different classification methods. Indeed, the SqueezeNet CNN performs slower than the other classification methods, with an average single-sample inference time of.0258 s. However, this inference time is not necessarily a concern for real-time applications of this technology, as violinists can realistically only perform a few (i.e., no more than a dozen) gestures per second; thus, the SqueezeNet CNN model will still be able to provide feedback to the user at a reasonable rate.

Interestingly, all the non-CNN models require a shorter inference time when using manual feature extraction than when using automated feature extraction, yet neither feature-extraction method yields consistently higher accuracies across the non-CNN models. The difference in inference times is likely due to the backbone CNN model extracting many more features to describe the same few features that are manually selected by the model engineer. In this study, 1000 automatically selected features from the backbone model achieve comparable model accuracies as the seven manually selected features. Manually extracted features have the advantage of being chosen using the engineer's expertise in the application domain, making the features more meaningful than the somewhat arbitrary features learned by the backbone CNN.

VI. RELIABILITY

To assess the robustness of radar-based feedback on violin bowing gestures, it is important to analyze factors that may impact the reliability of the collected gesture data. Here we analyze the impact of variations in the user-to-radar distance on the resulting time-Doppler data.

Practically, a violinist cannot situate themselves exactly 1 m from the radar upon every setup; there will inevitably be

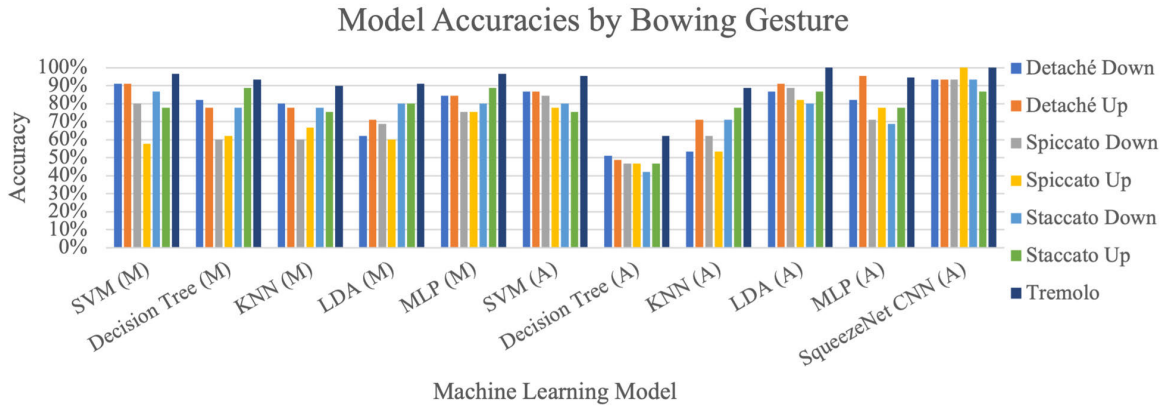


Fig. 7. Machine learning model accuracies for each type of bowing gesture. "(M)" indicates that the model used manually extracted features and "(A)" indicates that the model used automatically extracted features.

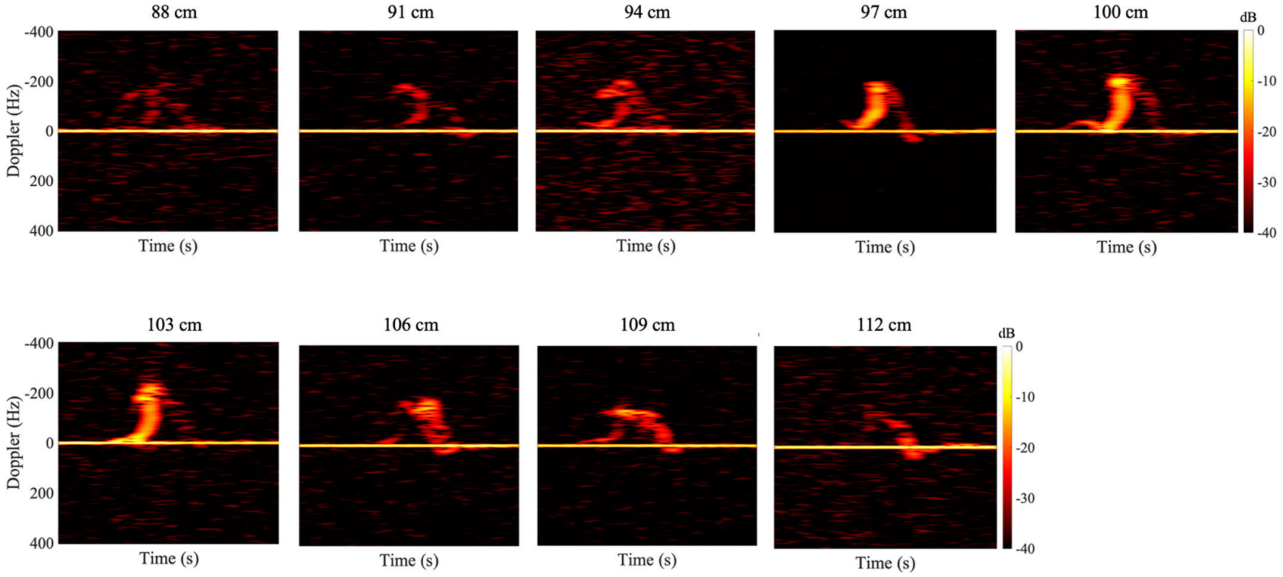


Fig. 8. Time-Doppler spectrograms for détaché down-bow gesture for different user-to-radar ranges.

TABLE V
VARIATIONS ON USER-TO-RADAR RANGE

Stroke	Minimum distance for a visible Doppler signature (cm)	Minimum distance for a clear Doppler signature (cm)	Maximum distance for a clear Doppler signature (cm)	Maximum distance for a visible Doppler signature (cm)
Détaché	88	91	109	112
Spiccato	91	94	109	112
Staccato	82	85	109	112
Tremolo	No minimum	No minimum	109	112

variation in the user's range between practices. Furthermore, in violin performance, it is common for the musician's body to shift its location, making it difficult for them to remain at precisely the same range from the radar throughout. These full-body movements may be an unintentional consequence of moving the bowing arm back and forth, or they may serve as an intentional conveyance of musical emotions [33], [34].

To analyze the impact of these user-to-radar distance variations on the time-Doppler spectrograms for various bowing motions, each of the down-bow strokes for each technique is performed at various ranges. First, each stroke is performed at the standard distance of 1 m from the radar; the violinist's distance for successive radar recordings is gradually decreased and increased in 3 cm intervals. Fig. 8 displays the time-Doppler spectrograms for the down-bow détaché gesture across different ranges. The minimum and maximum distances for producing a clear Doppler signature as well as the minimum and maximum distances for producing a visible Doppler signature are recorded. These values are displayed in Table V for all four bowing techniques.

Across all bowing techniques, the most conservative bounds for producing clear time-Doppler signatures are a minimum distance of 94 cm and a maximum distance of 109 cm. The most conservative bounds for producing detectable Doppler signatures are a minimum distance of 91 cm and a maximum distance of 112 cm. Thus, for reasonable gesture recognition,

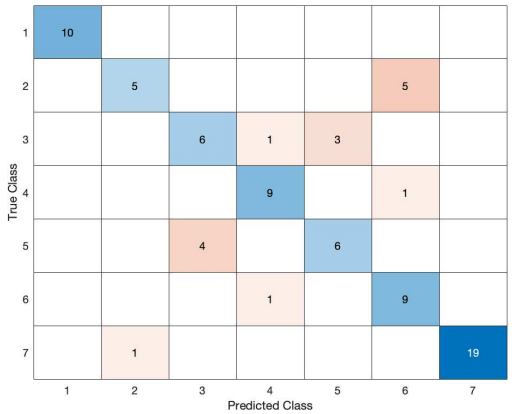


Fig. 9. Confusion matrix for the SqueezeNet CNN model's predictions on gestures recorded at the minimum and maximum distances necessary for a clear Doppler signature. Numbers 1–7 represent the classes in the following order: détaché down, détaché up, spiccato down, spiccato up, staccato down, staccato up, and tremolo.

the violinist has a 15 cm window of allowable ranges, which gives room for a modest amount of body swaying.

The SqueezeNet CNN model is used to evaluate the effect of spectrograms produced at different ranges on classification accuracy. Using the methods described in Section III, a total of 75 time-Doppler samples are recorded for all gestures performed at the minimum and maximum distances necessary for a clear Doppler signature (as reported in Table V). For the tremolo bow stroke, a minimum bound is arbitrarily set at a range of 31 cm, which provides more than sufficient room for large swaying motions from the user. A confusion matrix showing the SqueezeNet CNN model's predictions on this data can be found in Fig. 9. While some gestures, such as detaché down-bow and staccato up-bow, do not experience a decline in prediction accuracy at the minimum and maximum range bounds, other gestures do. The detaché up-bow suffers the most, dropping from an accuracy rate of 93%–50%, as the SqueezeNet CNN has trouble distinguishing the gesture from a staccato up-bow.

VII. Discussion and Conclusion

This study investigates the potential of using an FMCW radar to recognize various violin-bowing gestures. Previous studies have attempted to automate musical instrument feedback, but they relied on vision systems, sensor attachments to the violinist or the instrument, and wearable devices; these approaches pose problems such as privacy concerns, sensitivity to lighting, discomfort during the performance, and more. This work proposes the use of the radar, which does not suffer from those problems, as a method of recognizing bowing gestures by sensing the position and velocity of the bowing arm. Over a thousand samples of violin bowing gestures performed on the A string are collected from three subjects via radar. Through observing the corresponding time-Doppler spectrograms, the bowing gestures can be distinguished from one another due to differences in certain features such as maximum/minimum relevant frequency and time span. To automate bowing gesture classification, time-Doppler information for gestures is fed into machine learning classifiers. Both an automated and a manual approach to extracting features for training from the time-Doppler signature are explored. Ultimately, the fine-tuned pretrained SqueezeNet CNN model performs the best, with an accuracy of 95.00%. Although this model is outperformed marginally by the MLP (A) in classifying the detaché up-bow and the Decision Tree (M) and MLP (M) in classifying the staccato up-bow gesture, the SqueezeNet CNN most accurately predicts gestures across the board. An analysis of varying the user-to-radar range reveals that clear and visible time-Doppler spectrograms can be produced even in the presence of full-body swaying. However, if the range fluctuations are appreciable, the SqueezeNet CNN model struggles to accurately predict certain bowing gestures. Given the sensitivity of the machine learning model to significant range fluctuations, this technology is not suited for evaluating stagy performances but can reasonably be used for providing feedback in controlled practice settings.

As this study focuses on recognizing gestures only on the violin A string, future work should investigate the signatures for gestures performed on other violin strings. Changing between strings on the violin results in slight changes in the angle of the stroke relative to the radar, yielding unique signatures that must be learned by the classifiers. Furthermore, it may be worth analyzing micro-Doppler signatures and range profiles of bowing gestures to help isolate the desired hand-arm movements from extraneous movements and body swaying. Finally, as a pilot study for radar-based violin gesture recognition, the goal of this work was to explore the potential of radar and machine learning in differentiating the general forms of various bowing techniques. However, the ultimate goal of this technology is to not only identify the user's bowing gesture but also gauge how accurately the intended gesture was performed. Future work could explore the potential of radar and machine learning in providing a quantitative assessment of the deviation in gesture performance from the standard.

Acknowledgment

The authors would like to thank Prof. Annie Chalex Boyle, Rodrigo Cardona Cabrera, and Francisco Villarroel, all from Texas Tech University School of Music, for their support in performing the experiments. They would also like to acknowledge Christopher Williams and Victor G. Rizzi Varela, both from the Texas Tech University Department of Electrical and Computer Engineering, for their assistance in the initial setup of this project.

References

[1] C. Provenzale, N. D. Stefano, A. Noccaro, and F. Taffoni, "Assessing the bowing technique in violin beginners using MIMU and optical proximity sensors: A feasibility study," *Sensors*, vol. 21, no. 17, p. 5817, Aug. 2021, doi: 10.3390/s21175817.

[2] E. Schoonderwaldt and M. M. Wanderley, "Visualization of bowing gestures for feedback: The Hodgson plot," in *Proc. AXMEDIS*, Barcelona, Spain, 2007, pp. 65–70.

[3] A. Wilson, K. Hunter, and L. Moscardini, "Widening the gap? The challenges for equitable music education in Scotland," *Support Learn.*, vol. 35, no. 4, pp. 473–492, Oct. 2020, doi: 10.1111/1467-9604.12328.

[4] K. Ng and P. Nesi, "I-maestro: Technology-enhanced learning and teaching for music," in *Proc. NIME*, Genova, Italy, 2008, pp. 225–228.

[5] D. Young, "The hyperbow controller: Real-time dynamics measurement of violin performance," in *Proc. NIME*, Dublin, Ireland, 2002, pp. 1–6.

[6] P. Cook and D. Trueman, "BoSSA: The deconstructed violin reconstructed," *J. New Music Res.*, vol. 29, no. 2, pp. 121–130, Jun. 2000, doi: 10.1076/jnmr.29.2.121.3098.

[7] N. H. Rasamimanana, E. Fléty, and F. Bevilacqua, "Gesture analysis of violin bow strokes," in *Proc. Gesture Hum.-Comput. Interact. Simulation*, vol. 3881, 2006, pp. 145–155, doi: 10.1007/11678816_17.

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

[9] B.-Y. Liu, Y.-H. Jen, S.-W. Sun, L. Su, and P.-C. Chang, "Multimodal deep learning-based violin bowing action recognition," in *Proc. IEEE Int. Conf. Consum. Electron. Taiwan (ICCE-Taiwan)*, Sep. 2020, pp. 1–2, doi: 10.1109/ICCE-Taiwan49838.2020.9257995.

[10] E. Schoonderwaldt, N. Rasamimanana, and F. Bevilacqua, "Combining accelerometer and video camera: Reconstruction of bow velocity profiles," in *Proc. NIME*, Paris France, 2006, pp. 200–203.

[11] A. Perez-Carrillo and M. M. Wanderley, "Indirect acquisition of violin instrumental controls from audio signal with hidden Markov models," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 5, pp. 932–940, May 2015, doi: 10.1109/TASLP.2015.2410140.

[12] H. S. Alar, R. O. Mamaril, L. P. Villegas, and J. R. D. Cabarrubias, "Audio classification of violin bowing techniques: An aid for beginners," *Mach. Learn. with Appl.*, vol. 4, Jun. 2021, Art. no. 100028, doi: 10.1016/j.mlwa.2021.100028.

[13] D. Dalmazzo and R. Ramírez, "Bowing gestures classification in violin performance: A machine learning approach," *Frontiers Psychol.*, vol. 10, p. 344, Mar. 2019, doi: 10.3389/fpsyg.2019.00344.

[14] J. van der Linden, E. Schoonderwaldt, J. Bird, and R. Johnson, "MusicJacket—Combining motion capture and vibrotactile feedback to teach violin bowing," *IEEE Trans. Instrum. Meas.*, vol. 60, no. 1, pp. 104–113, Jan. 2011, doi: 10.1109/TIM.2010.2065770.

[15] D. Young, "Wireless sensor system for measurement of violin bowing parameters," in *Proc. SMAC*, Stockholm, Sweden, 2003, pp. 111–114.

[16] Y. Oyamada, T. Koshisaka, and T. Sakamoto, "Experimental demonstration of accurate noncontact measurement of arterial pulse wave displacements using 79-GHz array radar," *IEEE Sensors J.*, vol. 21, no. 7, pp. 9128–9137, Apr. 2021, doi: 10.1109/JSEN.2021.30526.

[17] G. Wang, J.-M. Muñoz-Ferreras, C. Gu, C. Li, and R. Gómez-García, "Application of linear-frequency-modulated continuous-wave (LFMCW) radars for tracking of vital signs," *IEEE Trans. Microw. Theory Techn.*, vol. 62, no. 6, pp. 1387–1399, Jun. 2014, doi: 10.1109/TMTT.2014.2320464.

[18] I. Choi, M. Kim, J. Choi, J. Park, S. Park, and K. Kim, "Robust cardiac rate estimation of an individual," *IEEE Sensors J.*, vol. 21, no. 13, pp. 15053–15064, Jul. 2021, doi: 10.1109/JSEN.2021.3074510.

[19] Y. Li, Z. Peng, R. Pal, and C. Li, "Potential active shooter detection based on radar micro-Doppler and range-Doppler analysis using artificial neural network," *IEEE Sensors J.*, vol. 19, no. 3, pp. 1052–1063, Feb. 2019, doi: 10.1109/JSEN.2018.2879223.

[20] L. Gan, Y. Liu, Y. Li, R. Zhang, L. Huang, and C. Shi, "Gesture recognition system using 24 GHz FMCW radar sensor realized on real-time edge computing platform," *IEEE Sensors J.*, vol. 22, no. 9, pp. 8904–8914, May 2022, doi: 10.1109/JSEN.2022.3163449.

[21] S. Zhang, G. Li, M. Ritchie, F. Fioranelli, and H. Griffiths, "Dynamic hand gesture classification based on radar micro-Doppler signatures," in *Proc. CIE Int. Conf. Radar (RADAR)*, Oct. 2016, pp. 1–4, doi: 10.1109/RADAR.2016.8059518.

[22] H. Gao, C. Williams, V. G. R. Varela, and C. Li, "Violin gesture recognition using FMCW radars," in *Proc. IEEE Top. Conf. Wireless Sensors Sensor Netw.*, Las Vegas, NV, USA, 2023, pp. 13–15, doi: 10.1109/WiSNeT56959.2023.10046213.

[23] Z. Peng et al., "A portable FMCW interferometry radar with programmable low-IF architecture for localization, ISAR imaging, and vital sign tracking," *IEEE Trans. Microw. Theory Techn.*, vol. 65, no. 4, pp. 1334–1344, Apr. 2017, doi: 10.1109/TMTT.2016.2633352.

[24] F. Shaheen, B. Verma, and M. Asafuddoula, "Impact of automatic feature extraction in deep learning architecture," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2016, pp. 1–8, doi: 10.1109/DICTA.2016.7797053.

[25] A. Pradhan, "Support vector machine—A survey," *Int. J. Emerg. Technol.*, vol. 2, no. 8, pp. 82–85, Aug. 2020.

[26] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *J. Chemometrics*, vol. 18, no. 6, pp. 275–285, Jun. 2004, doi: 10.1002/cem.873.

[27] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.

[28] H. Ramchoun, M. A. J. Idrissi, Y. Ghanou, and M. Ettaouil, "Multilayer perceptron: Architecture optimization and training," *Int. J. Interact. Multi*, vol. 4, no. 1, pp. 26–30, Jan. 2016, doi: 10.9781/ijimai.2016.415.

[29] K. Liu, G. Kang, N. Zhang, and B. Hou, "Breast cancer classification based on fully-connected layer first convolutional neural networks," *IEEE Access*, vol. 6, pp. 23722–23732, 2018, doi: 10.1109/ACCESS.2018.2817593.

[30] C. Ding et al., "Inattentive driving behavior detection based on portable FMCW radar," *IEEE Trans. Microw. Theory Techn.*, vol. 67, no. 10, pp. 4031–4041, Oct. 2019, doi: 10.1109/TMTT.2019.2934413.

[31] S. Skaria, A. Al-Hourani, M. Lech, and R. J. Evans, "Hand-gesture recognition using two-antenna Doppler radar with deep convolutional neural networks," *IEEE Sensors J.*, vol. 19, no. 8, pp. 3041–3048, Apr. 2019, doi: 10.1109/JSEN.2019.2892073.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[33] S. Dahl and A. Friberg, "Visual perception of expressiveness in Musicians' body movements," *Music Perception*, vol. 24, no. 5, pp. 433–454, Jun. 2007, doi: 10.1525/mp.2007.24.5.433.

[34] J.-W. Liu, H.-Y. Lin, Y.-F. Huang, H.-K. Kao, and L. Su, "Body movement generation for expressive violin performance applying neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3787–3791, doi: 10.1109/ICASSP40776.2020.9054463.

Hannah Gao is currently a Senior with the Harriton High School, Rosemont, PA, USA. As a part of the 2023 Clark Scholars Summer Research Program, she conducted research in electrical and computer engineering at Texas Tech University, Lubbock, TX, USA. Her current research interests include radar technology and machine learning.

Changzhi Li (Senior Member, IEEE) received the B.S. degree in electrical engineering from Zhejiang University, Hangzhou, China, in 2004, and the Ph.D. degree in electrical engineering from the University of Florida, Gainesville, Florida, in 2009.

He is a Professor at Texas Tech University, Lubbock, TX, USA. His research interests include microwave/millimeter-wave sensing for healthcare, security, energy efficiency, structural monitoring, and human–machine interface.

Dr. Li is a Fellow of the National Academy of Inventors (NAI). He is an IEEE Microwave Theory and Techniques Society (MTT-S) Distinguished Microwave Lecturer. He was a recipient of the IEEE Sensors Council Early Career Technical Achievement Award.