MACHINE LEARNING-BASED CLASSIFICATION OF GUITAR PLAYING TECHNIQUES

Ziyang Ou, Xiaohe Tian, Jiayin Yuan

Department of Electrical and Computer Engineering, University of Rochester, Rochester, New York, USA

ABSTRACT

Improvisation is often found in jazz and blues music, and it is not typically notated down. It would be helpful to music coaching and instruction if the guitar player's technique or note could be recognized and marked down automatically. However, at this point, the detection of performance techniques is still at an early stage, and only a limited number of techniques can be detected. This is partly because there are not many training video materials, and partly because audio data contains a lot of information, audio patterns spread sparsely in different scenarios. It is challenging to generalize a common classification model. The aim of this experiment is to verify the viability of automatic electric guitar playing techniques recognition and compare the performance of various feature extraction and audio data augmentation techniques. We collected a dataset of nine techniques, two feature-extraction pipelines were compared: (1) wav2vec model + MLP, a transfer learning method extending speech representation embeddings to instrumental sounds; (2) Mel-spectrogram + -ConvNet, a computer vision--based pipeline. Results showed Top-1 accuracy was (1) 65%, (2) 93% respectively. Experiments have shown that self-supervised speech models can be migrated to another sound type recognition task without retraining from scratch. And traditional ConvNet architecture performed well in sound recognition tasks. We also conducted data augmentation to enhance model's generalization ability, allowing models to adapt to different recording conditions, playing styles or tonal variations.
Results demonstrate that guitar technique recognition is feasible and promising for music education, transcription, and performance analysis applications.

Index Terms— Machine Learning, Music Information Retrieval, Audio Signal Processing

1. INTRODUCTION

"Music is the universal language of mankind," Henry Wadsworth Longfellow wrote. Much like spoken language, which not only uses words to convey meaning but also the way it is delivered – by shouting, whispering or crying. Music conveys emotion and expression through performance techniques. In the case of guitar, there are techniques such as bending, sliding, vibrato, or tapping. They were used as expressive devices, adding additional m eanings to a note. In genres like jazz and blues music where improvisation often appears and is seldom notated in scores. Yet, this kind of music is often left un-transcribed, differing from every performer, making it difficult to analyze, teach, or preserve systematically.
Research on music information retrieval has made progress on note transcription, drumbeat detection, and instrument classification, the recognition of instrumental performance techniques is still developing. Most previous work focuses on instruments like drums or piano, which their sound patterns are more obvious for detection. In contrast, the guitar presents unique challenges: the same note can be played using different techniques, and techniques may vary subtly across styles, instruments, and players. This task is further complicated by the nature of audio data. Performance techniques varied in tiny length of time spans and spectral variations, which are often obscured by noise, reverberation, or overlapping effects in real-world recordings, making it difficult for traditional classification models to generalize.
In this study, we aim to evaluate the feasibility of recognition based on machine learning of electric guitar playing techniques. Specifically, we investigate into two approaches to feature extraction and classification: (1) a transfer learning method using the model wav2vec followed by a lightweight multilayer perceptron (MLP) classifier, and (2) Mel-spectrograms as input to ConvNet. To improve model's generalization ability under varying recording and playing conditions, we apply several audio data augmentation operations on the dataset, e.g., pitch shifting, gain adjust, and additive noise. The dataset used in this study includes nine guitar playing techniques, preprocessed as fixed-length waveform segments as input. Our results show that the Mel spectrogram + ConvNet approach outperforms the wav2vec + MLP method, vision-based method achieves a Top-1 accuracy of 93% compared to 65%. Nevertheless, both models demonstrate that guitar playing technique recognition is viable.

2. RELATED WORK

Early studies focused on feature extraction engineering combined with classical machine learning models to classify isolated notes, while more recent work incorporates temporal modeling and multimodal approaches.

One of early feature-based detection method was conducted by Abeßer et al. (2010) [1] on electric bass guitar. They introduced the Bass dataset, covering ten common plucking and expression techniques, and used spectral and envelope-based features alongside support vector machines (SVMs), improved the classification accuracy by over 27% points in comparison to a state-of-the-art baseline system.

Su et al. (2014) [2] proposed a method for classifying seven electric guitar techniques. Their results demonstrated that combining spectral and phase information significantly improved the classification of electric guitar techniques.

For a complete transcription of a guitar instead of single note classification, Chen et al. [3] designed a two-stage framework for detecting lead guitar techniques in full-length solo recordings. They first identify prominent candidates by analyzing the extracted melody contour and then apply a pre-trained classifier to the candidates for playing technique detection using a set of timbre and pitch features. With combination, their approach showed strong performance on continuous solo phrases.

Violin shares the same sound-logic as many string instruments as guitar. Dalmazzo and Ramírez (2019) [4] proposed bowing gesture recognition in violin performance. After extracting features from both the motion and audio data, we trained an HHMM to identify the different bowing techniques automatically. Their model can determine the studied bowing techniques with over 94% accuracy.

In sum, these studies demonstrate both classic and modern machine learning architecture in recognizing expressive playing techniques. Key challenges include audio noise effects, performing habits from different players, and the need for high-quality labeled datasets.

3. METHODS

3.1. Dataset and Data augmentation

This study uses a publicly available dataset introduced by Pedroza et al. (2023) [5], designed for automatic recognition of electric guitar playing techniques. The dataset includes nine guitar techniques and below are detailed descriptions for each technique:

Alternate picking: continuously alternating between downward and upward strokes of the guitar strings using a pick.

Legato: playing the notes in a smooth manner, transitioning from one note to the next without leaving any audible space in between.

Tapping: using the fingertips of both the fretting and picking hand on the fretboard to fret notes.

Sweep Picking: using a 'sweeping' motion of the pick across strings resulting in a fast, fluid, and smooth sound.

Vibrato: repeatedly moving a string back and forth using the fretting hand, while holding a note, causing the pitch of the note to fluctuate.

Hammer-on: playing a note and then sharply bringing down a finger of the fretting hand onto the same string but on a higher fret from the first note played, without re-plucking the string.

Pull - off: removing a finger of the string-pressing hand from a previously played note, allowing the tone to ring and produce a lower tone.

Slide: smoothly moving the fretting hand along a string, across multiple frets, while keeping the string pressed down, resulting in a transition between notes.

Bend: stretching a string with the fretting hand after a note has been played, to increase the pitch of the played note.

Offline augmentation method: while the original dataset provides a diverse range of techniques, the distribution of recordings across classes is unbalanced. We applied data augmentation, pitch shift method specifically, to increase the number of training samples for each class, after adding roughly 4,500 samples in total (900 recordings × 25 sec duration ÷ 5sec for each sample duration).

On-the-fly augmentation method: Input data will be transformed dynamically as each training batch is loaded, actions like gain altering, random white noise adding.

3.2. Wav2vec

Wav2vec is a self-supervised speech model originally designed for speech recognition [6], but its capability to extract rich audio representations makes it well-suited for other audio classification tasks, such as guitar technique recognition. In our approach, raw guitar-playing waveforms are first passed through the pretrained wav2vec model to extract meaningful features. These features are then processed by a lightweight classification head composed of three fully connected layers.

Each of the first two layers is followed by a GELU activation and a dropout layer. The final output is a nine-dimensional vector representing the probability distribution over nine guitar technique classes. The predicted class is the one with the highest probability. Fig. 1 illustrates the end-to-end model pipeline.

Notably, since the embeddings produced by the wav2vec model inherently retain temporal information, we segmented the input audio into ~5-second clips to leverage wav2vec's strength in handling long audio sequences.

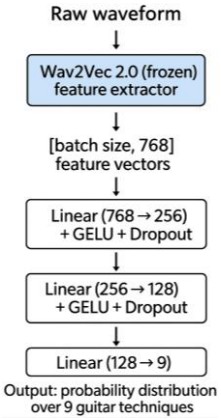

Fig. 1: Transferring learning framework

3.3. Mel-spectrogram + ConvNet

While we consider Wav2vec to be a reliable model for feature extraction in guitar technique classification, we would like to implement a feature extraction model from scratch, so that the model can be more specific to the task of guitar technique classification. By doing this, we first introduce Mel-spectrogram, a three-dimensional representation of audio energy intensity versus frequency and time. The frequency axis in Mel-spectrogram is obtained with a nonlinear transformation, from the real frequency scale into the frequency scale adjusted for human perception.

From the waveforms of guitar playing audio, we get the Mel-spectrogram patterns for the nine types of guitar playing techniques. We can see from Fig. 2 that the Mel-spectrogram patterns for different guitar playing techniques have distinct features, such as bright vertical lines for "hammer on" and "pull off" that imply energy bursts at times, and multiple horizontal lines for "bend" and "slide" that imply changes in pitch with time. Therefore, it is feasible to use the Mel-spectrogram data as the input for feature extraction.

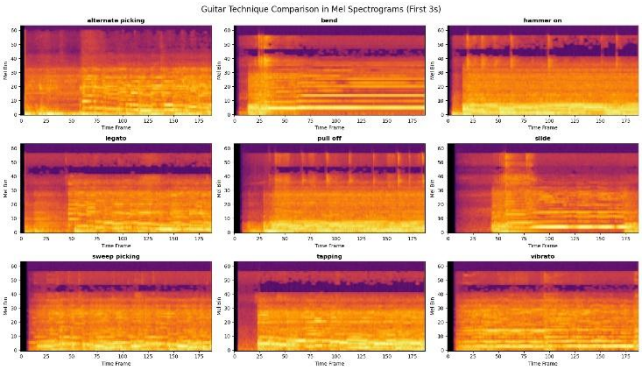

Fig. 2: Mel-spectrogram pattern of all 9 techniques.

After taking Mel-spectrogram transformation for input audio waveform data, we feed the Mel-spectrogram data into a convolutional neural network (ConvNet) for feature extraction. This feature extraction model is structured as such: we first convert the Mel-spectrogram data to decibels, and then apply a two-dimensional batch normalization to the converted data. Next, we apply four convolutional layers and add flatten layer action.

For guitar technique classification, we take the obtained feature extraction data and feed them into three fully connected layers, each of the first two layers followed by a leakyReLU activation function. Finally, we get the output, an array with a length of 9, representing the probability distribution for the nine guitar technique labels. The label with the highest probability is therefore the predicted output. Fig. 3 shows the complete pipeline of the Mel-spectrogram + ConvNet learning model.

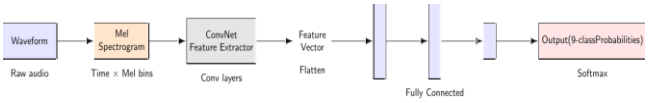

Fig. 3: Vision-based learning framework.

4. EXPERIMENTS

4.1. Training process

Training, validation and testing set are split in a ratio of 6:2:2. We use cross-entropy loss as the loss criterion and AdamW (with lr = 1e-4) as the optimizer for both models. Batch size for both methods is set to 64. We use a learning rate reduction scheme to dynamically reduce the learning rate when the improvement on metrics is low. We also implement an early stopping mechanism to stop training after the validation accuracy does not reach the best accuracy for 5 epochs consecutively.

4.1.1. Wav2vec route

Early stopped at 19-th epoch.

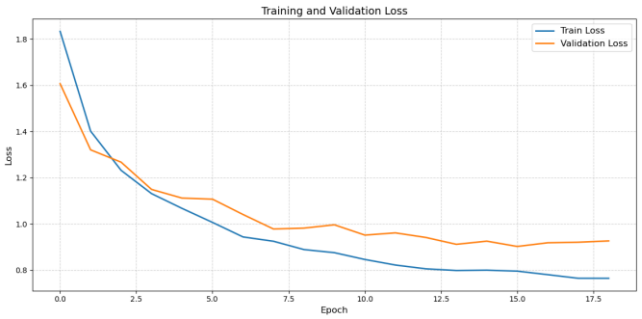

Fig. 4: Train and validation loss curve.

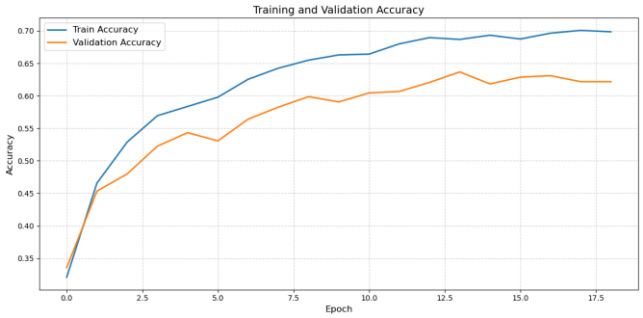

Fig. 5: Train and validation accuracy curve.

4.1.2. ConvNet route
Early stopped at 26-th epoch.

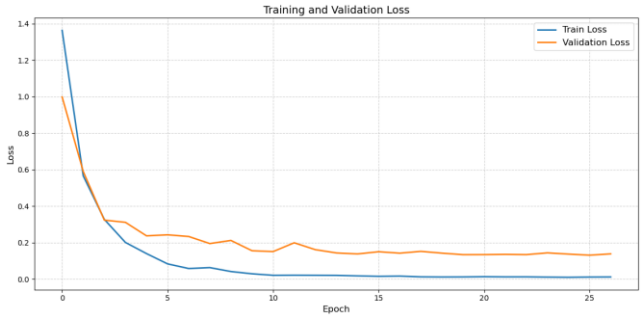

Fig. 6: Train and validation loss curve.

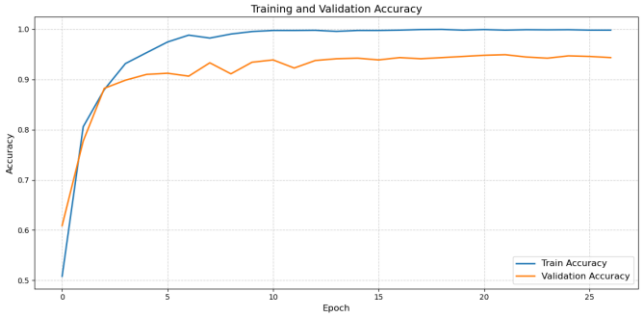

Fig. 7: Train and validation loss curve.

4.2. Performance on test set
We use three metrics to evaluate the performance: precision, recall, and F1-score.

4.2.1. Wav2vec route
Confusion matrix is shown on Fig. 8. The values of performance metrics corresponding to every label are shown in Fig. 9.

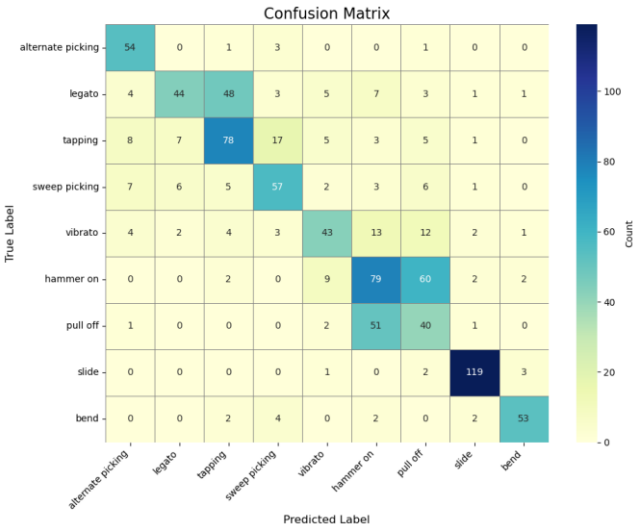

Fig. 8: confusion matrix of wav2vec model.

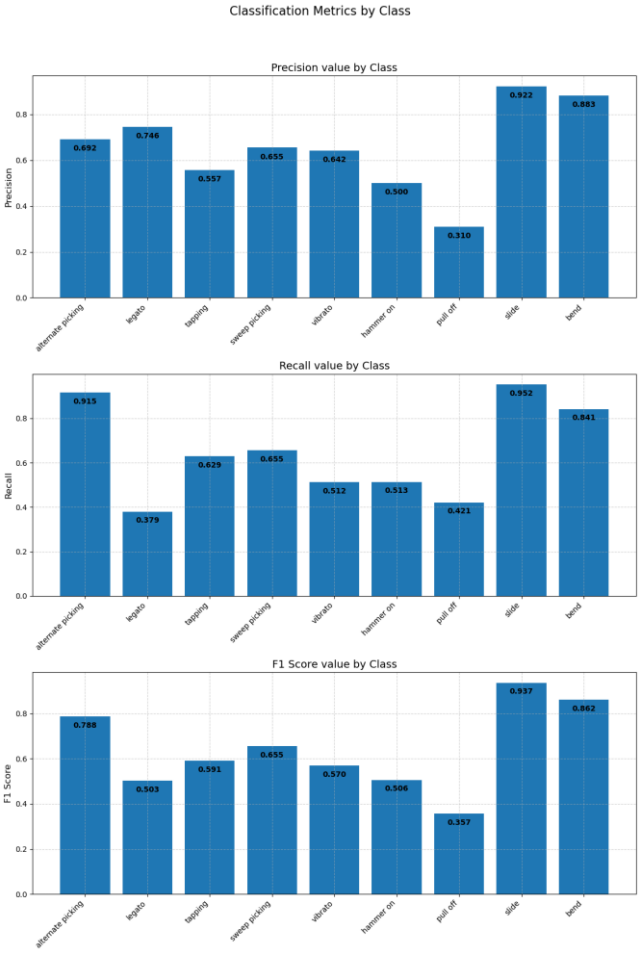

Fig. 9: performance by labels of wav2vec model.

4.2.2. ConvNet route

Confusion matrix is shown on Fig. 10. The values of performance metrics corresponding to every label are shown in Fig. 11.

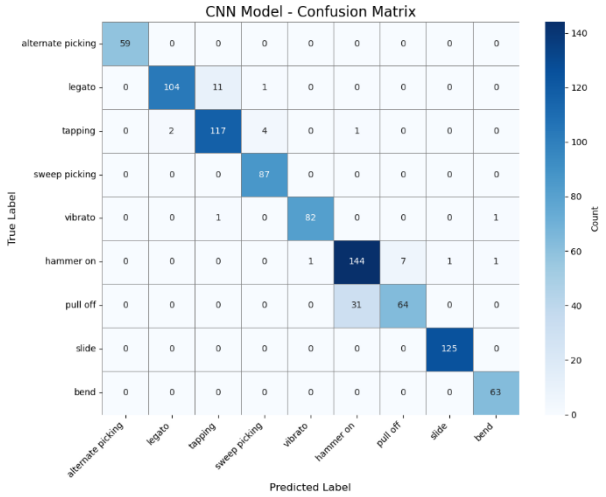

Fig. 10: Confusion matrix of ConvNet model.

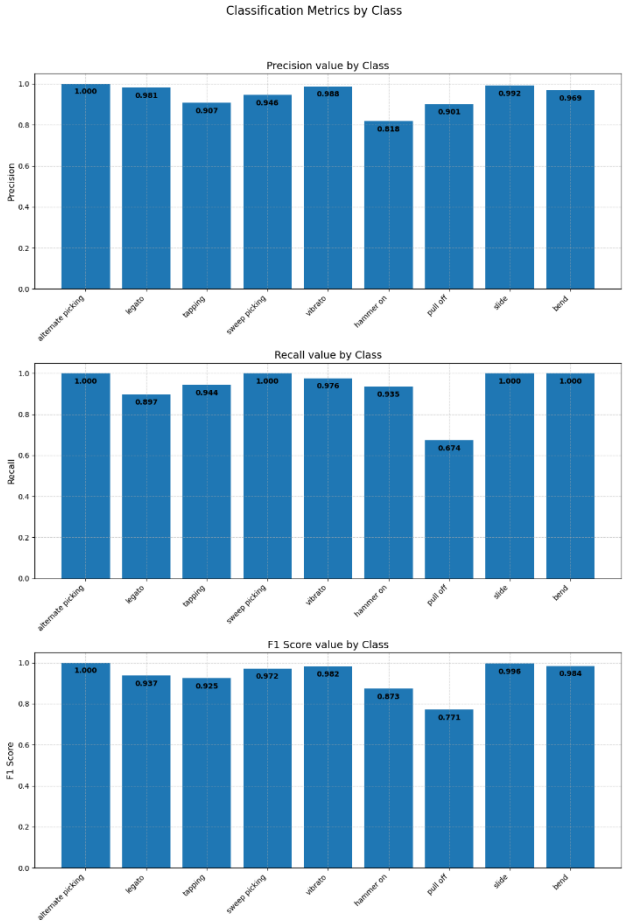

Fig. 11: performance by labels of ConvNet model.

4.3. Analyses of Results

A major issue emerged from both models is the mixing up of "hammer on" and "pull off" labels. We can see from the figures that a significant number of samples of both labels are misclassified as the other, resulting in lower performance for the two labels in all three metrics. There is also an issue in misclassifying "legato" as "tapping", resulting in the lowered performance of recall for the former, and precision for the latter. Since we can detect the different features between those labels in Mel-spectrogram patterns, there is still room for improvement for the ConvNet model.

In general, however, we can see that the ConvNet model has better performance than wav2vec model on all metrics and most labels. This observation indicates that the idea works that we use a guitar playing-specific feature extraction model to replace wav2vec, and Mel-spectrogram is a feasible medium for implementing this model.

4.4. Feature Space Visualization Before Classification

An interesting investigation is whether the extracted features are discriminative before going to the classifiers. This is a good exploration to see if the model can "hear" the differences instead of just "remember" samples from training set. We apply UMAP (Uniform Manifold Approximation and Projection) to visualize the high-dimensional feature embeddings on the validation set.

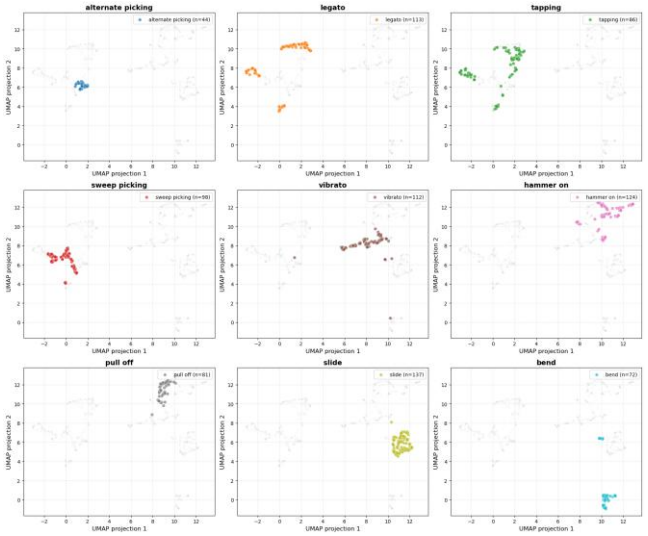

Fig. 12: UMAP projection graph of feature embeddings on the validation set.

As shown in the graph above, it can be observed that certain guitar techniques such as alternate picking, tapping, and bend form clear and well-separated clusters. However, several overlapping regions can be observed among techniques like hammer-on, pull-off, and vibrato, which are acoustically similar and often confused due to their shared temporal and spectral patterns. This overlap in the 2D projection does not necessarily imply that these classes are

inseparable. The features have been compressed from a high-dimensional space to a two-dimensional plane just for visualization purposes, and it is possible that decision boundaries exist in the original high-dimensional latent space.

5. CONCLUSION AND FUTURE WORK

This study investigated the feasibility of automatic electric-guitar playing-technique recognition using two pipelines: (1) wav2vec 2.0 + MLP (transfer learning) – a powerful self-supervised pre-trained network mainly on speech feature representation model adapt on musical audio. (2) ConvNetwork – an end-to-end network trained from scratch. To address the class imbalance problem and make the results of the experiment more convincing, we performed data augmentation on the origin dataset. Pitch-shift augmentation was performed to equalize each of nine techniques, after it the whole dataset comes to 4500 samples in total. On-the-fly augmentations (gain variation, additive noise) further promoted robustness during training. Experimental results show that the vision-based pipeline achieved a Top-1 accuracy of 93%, outperformed the transfer-learning route which is 65%. UMAP visualization method confirmed that distinct playing techniques clusters are already formed in the feature space before classification.

In summary, our experiments show that: (1) guitar techniques can be recognized via time-frequency graph representation. (2) transfer learning route is feasible, speech-based self-supervised models offer useful feature extractors for other audio-classification tasks. (3) a balanced and augmented dataset is important for a reliable result.

Future work could involve (1) exploring other combinations of hyper-parameter settings and network architectures to further improve accuracy. (2) extract embeddings from different large pre-trained network beyond speech-based Wav2vec model. (3) curating additional datasets with diverse recording setup: different players, guitars, amplifiers, and microphone, to test generalization. (4) applying the framework to other string instruments like violin, bass, and cello.

6. REFERENCE

[1] J. Abesser, H. Lukashevich, and G. Schuller, "Feature-based extraction of plucking and expression styles of the electric bass guitar," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, TX: IEEE, Mar. 2010, pp. 2290–2293. doi: 10.1109/ICASSP.2010.5495945.

[2] L. Su, L.-F. Yu, and Y.-H. Yang, "SPARSE CEPSTRAL AND PHASE CODES FOR GUITAR PLAYING TECHNIQUE CLASSIFICATION," 2014.

[3] Y.-P. Chen, L. Su, and Y.-H. Yang, "ELECTRIC GUITAR PLAYING TECHNIQUE DETECTION IN REAL-WORLD RECORDINGS BASED ON F0 SEQUENCE PATTERN RECOGNITION".

[4] D. Dalmazzo and R. Ramírez, "Bowing Gestures Classification in Violin Performance: A Machine Learning Approach," *Front. Psychol.*, vol. 10, p. 344, Mar. 2019, doi: 10.3389/fpsyg.2019.00344.

[5] A. Mitsou, A. Petrogianni, E. A. Vakalaki, C. Nikou, T. Psallidas, and T. Giannakopoulos, "A multimodal dataset for electric guitar playing technique recognition," *Data Brief*, vol. 52, p. 109842, Feb. 2024, doi: 10.1016/j.dib.2023.109842.

[6] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Oct. 22, 2020, *arXiv*: arXiv:2006.11477. doi: 10.48550/arXiv.2006.11477.

[7] A. van den Oord *et al.*, "WaveNet: A Generative Model for Raw Audio," Sep. 19, 2016, *arXiv*: arXiv:1609.03499. doi: 10.48550/arXiv.1609.03499.

[8] M. I. Ansari and T. Hasan, "SpectNet : End-to-End Audio Signal Classification Using Learnable Spectrograms," Nov. 17, 2022, *arXiv*: arXiv:2211.09352. doi: 10.48550/arXiv.2211.09352.