Guitar Technique Classification

Ziyang Ou

Jiayin Yuan

Xiaohe Tian

ECE 408 The Art of Machine Learning

Department of Electrical and Computer Engineering

Background and motivation

- Guitar playing technique, which covers from basic strumming to advanced picking, is a way for guitarists to express musical ideas and emotions.

- Some examples:

Bend

Pull Off

Vibrato

One who has never played the guitar can tell the difference through hearing.

Background and motivation

- Guitar playing technique has strong acoustical characteristic, would it be possible to use machine to learn these acoustic patterns?
  - Answer: Yes

- Purposed two “learning” routes:
  - 1. Mel Spectrogram + Convolutional Neural Network (ConvNet)

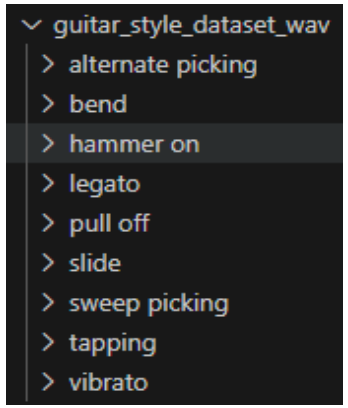
  - 2. Transfer Learning based on wav2vec-2.0

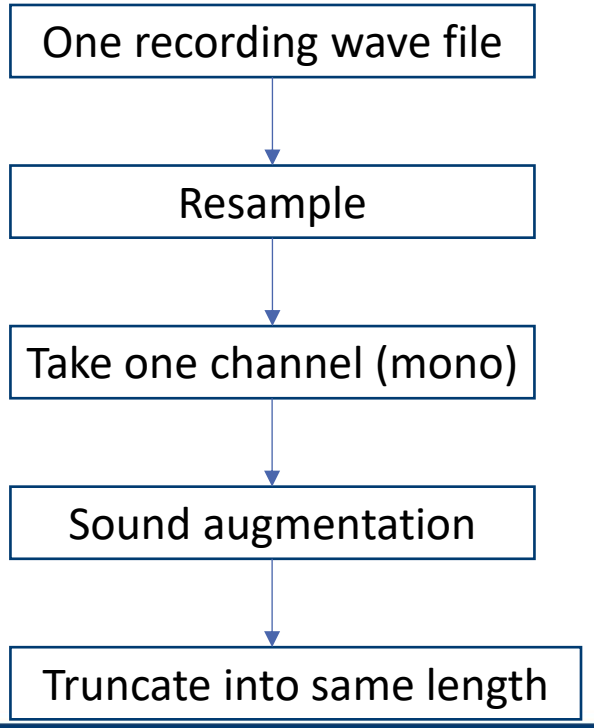
Dataset and data preprocessing

- 9 types of guitar techniques are labeled in the dataset. Each sample is stored as .wav files with 16bits 48kHz sample rate.

- DataSet in torch library, provides one easy way to pre-process the data.

- Data augmentation





Preprocessing process

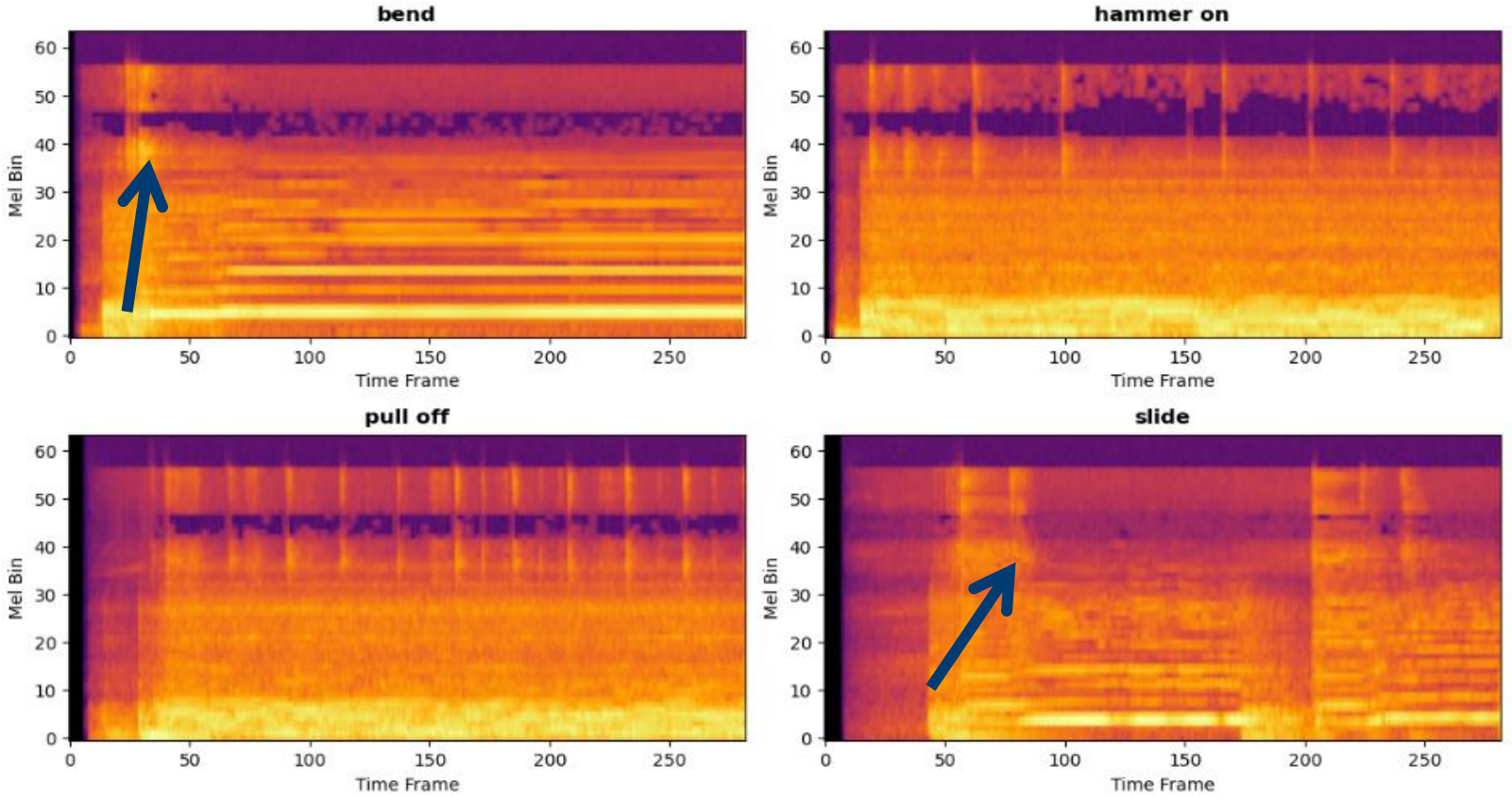
Learning Framework: Mel Spectrogram + ConvNet

- Mel Spectrogram: representation of short-term power spectrum of sound
  - x-axis: Time
  - y-axis: Mel frequency bin(unlike traditional spectrogram, Mel frequency is not in linear scaling, but in nonlinear scaling based on human perception), low-frequency portion is less spaced.
  - brightness: Amplitude

- Hammer on/pull off: bright lines at high-frequencies

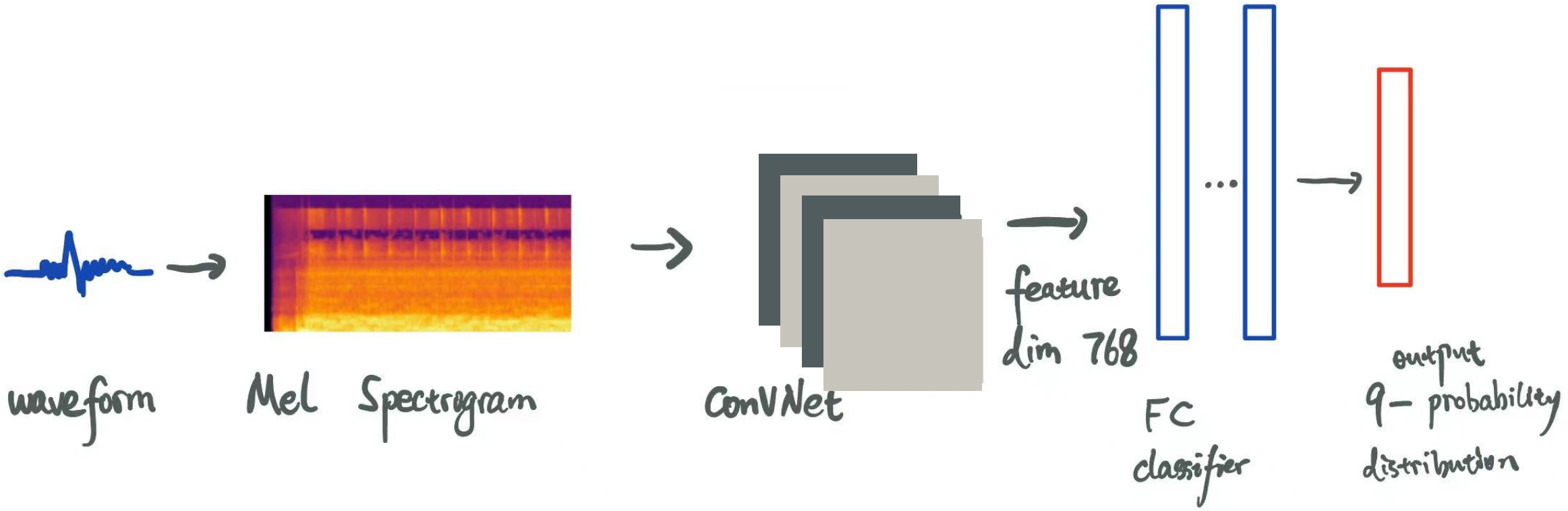
- Bend: change in a short time period

- Slide: change from low to high frequency in 0.5sec

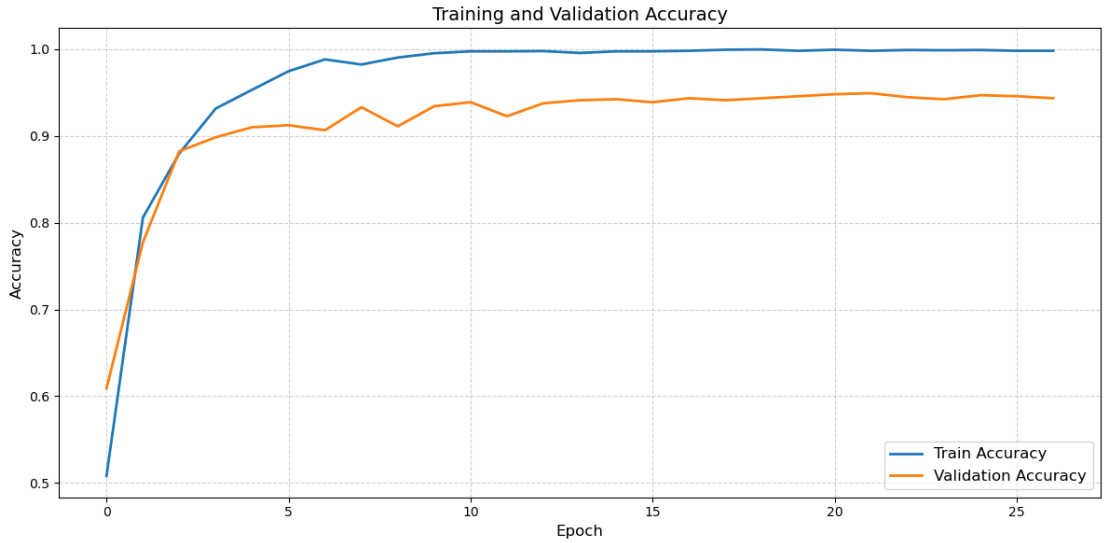


Learning Framework: Mel Spectrogram + ConvNet

Neural Network Workflow Diagram



Learning Framework: Mel Spectrogram + ConvNet




```
optimizer = torch.optim.Adam(cnn_model.parameters(), lr=1e-4, weight_decay=5e-4)
criterion = nn.CrossEntropyLoss()
```

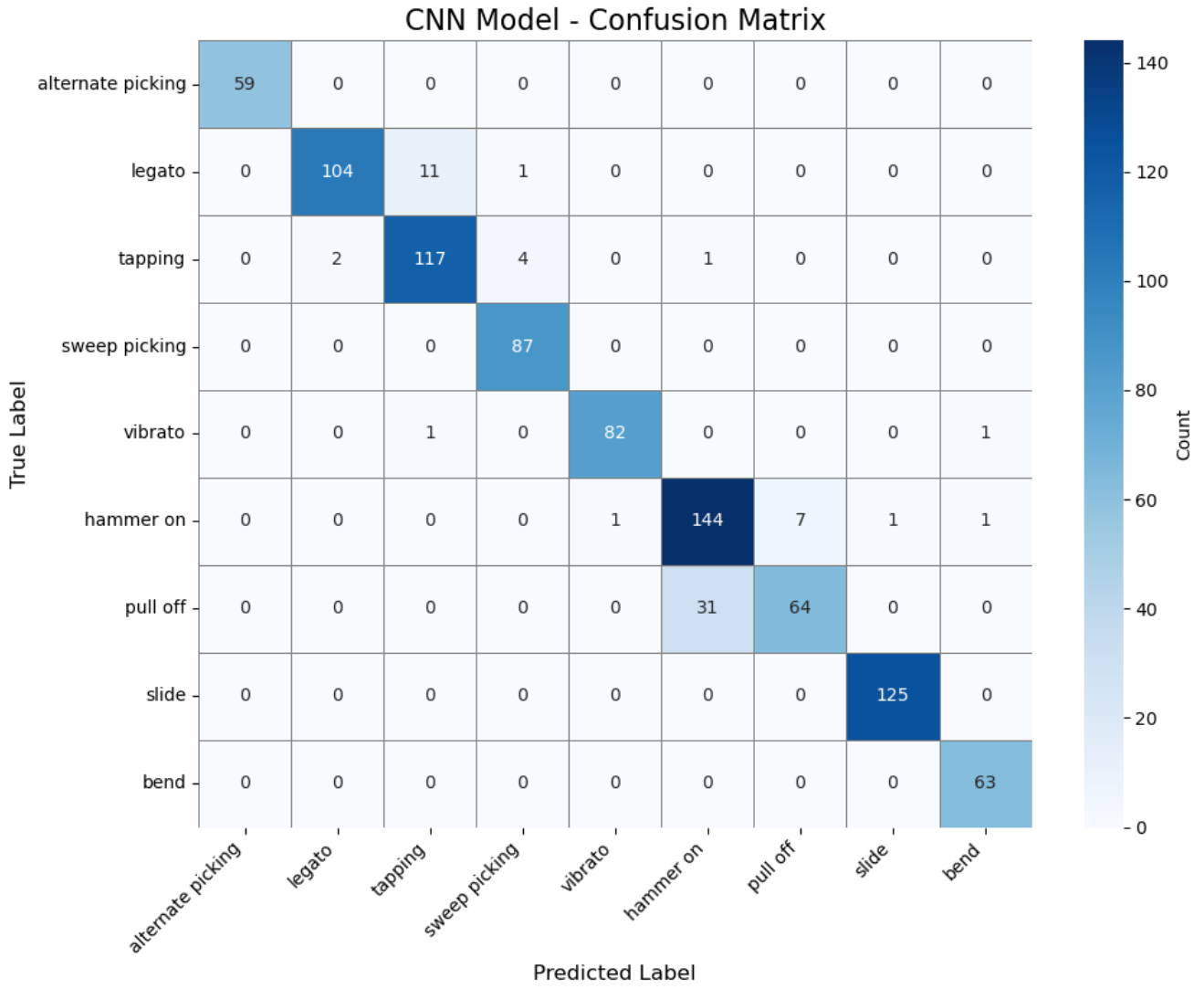
With early stopping mechanism, stops at around 27-th epochs

Learning Framework: Mel Spectrogram + ConvNet

Classification Report:

	precision	recall	f1-score	support
alternate picking	1.00	1.00	1.00	59
legato	0.98	0.90	0.94	116
tapping	0.91	0.94	0.92	124
sweep picking	0.95	1.00	0.97	87
vibrato	0.99	0.98	0.98	84
hammer on	0.82	0.94	0.87	154
pull off	0.90	0.67	0.77	95
slide	0.99	1.00	1.00	125
bend	0.97	1.00	0.98	63
accuracy			0.93	907
macro avg	0.94	0.94	0.94	907
weighted avg	0.93	0.93	0.93	907

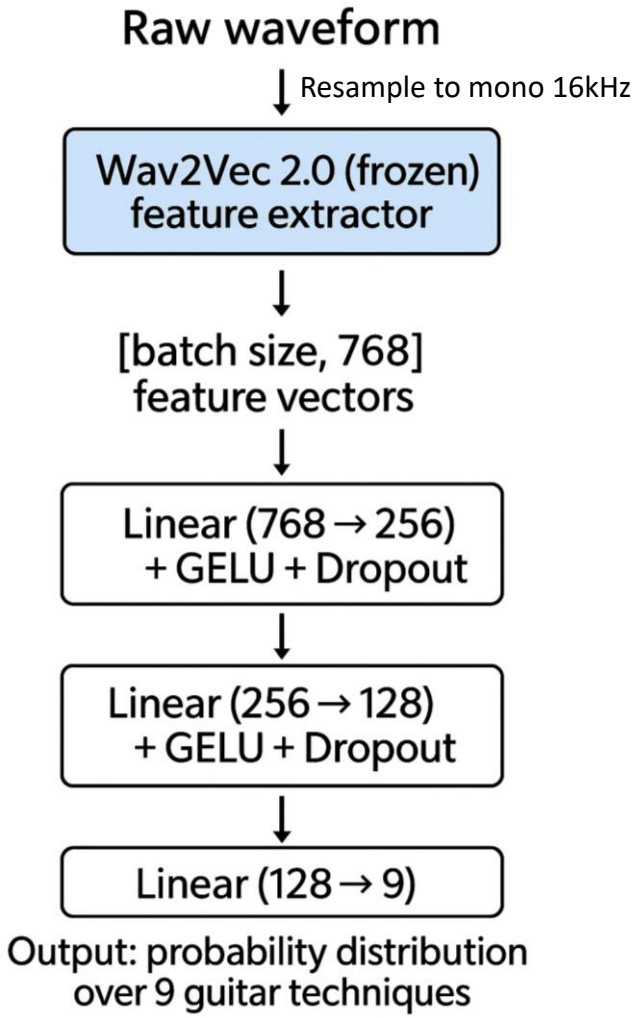
- The model can “see” from spectra and tell the difference between classes.
- Results are amazingly good, even hammer on (lower note appear first), pull off (higher note appear first) can be classified.



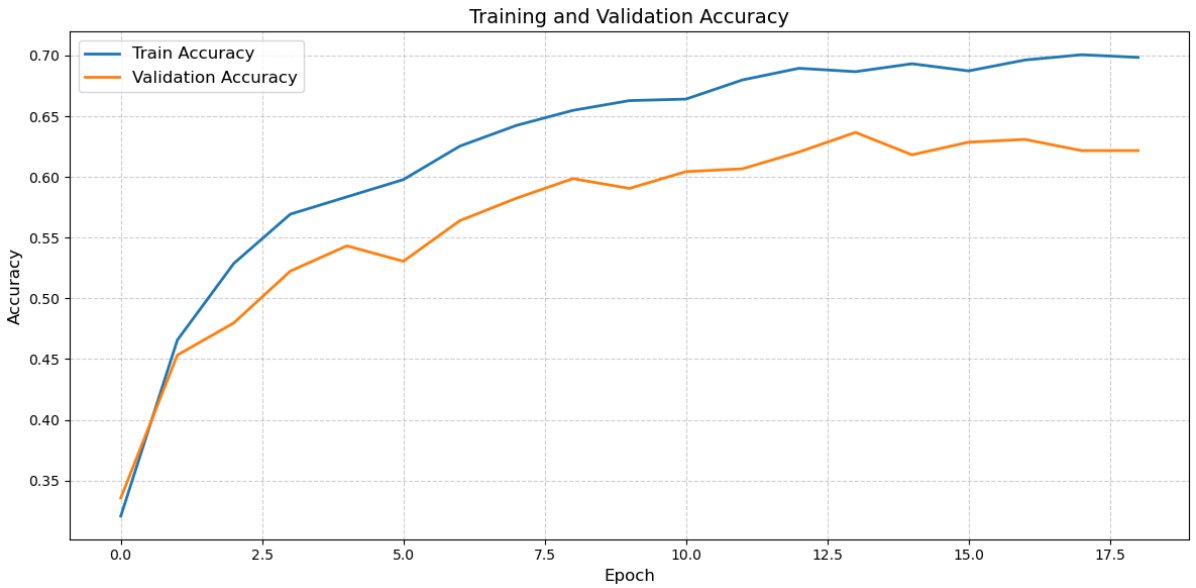
Learning Framework: Wav2Vec 2.0 + linear classifier

- Wav2vec-2.0 (Facebook) is a framework used self-supervised learning method on speech representations, that is, by pre-training on large amounts of unlabeled speech data so the model can capture the acoustic and linguistic features of speech.

- Although this model trains on large amounts of unlabeled **speech** data, the audio properties of guitar may share some commons with speech.
- Idea: transfer and retrain a new boundary for guitars Worth a try!



Learning Framework: Wav2Vec 2.0 + linear classifier



With early stopping mechanism, stops at around 20-th epochs

It is understandable that WavNet2.0 structure may not as well as the previous (computer-vision based) one, because few guitar sound was trained on Wav2vec.

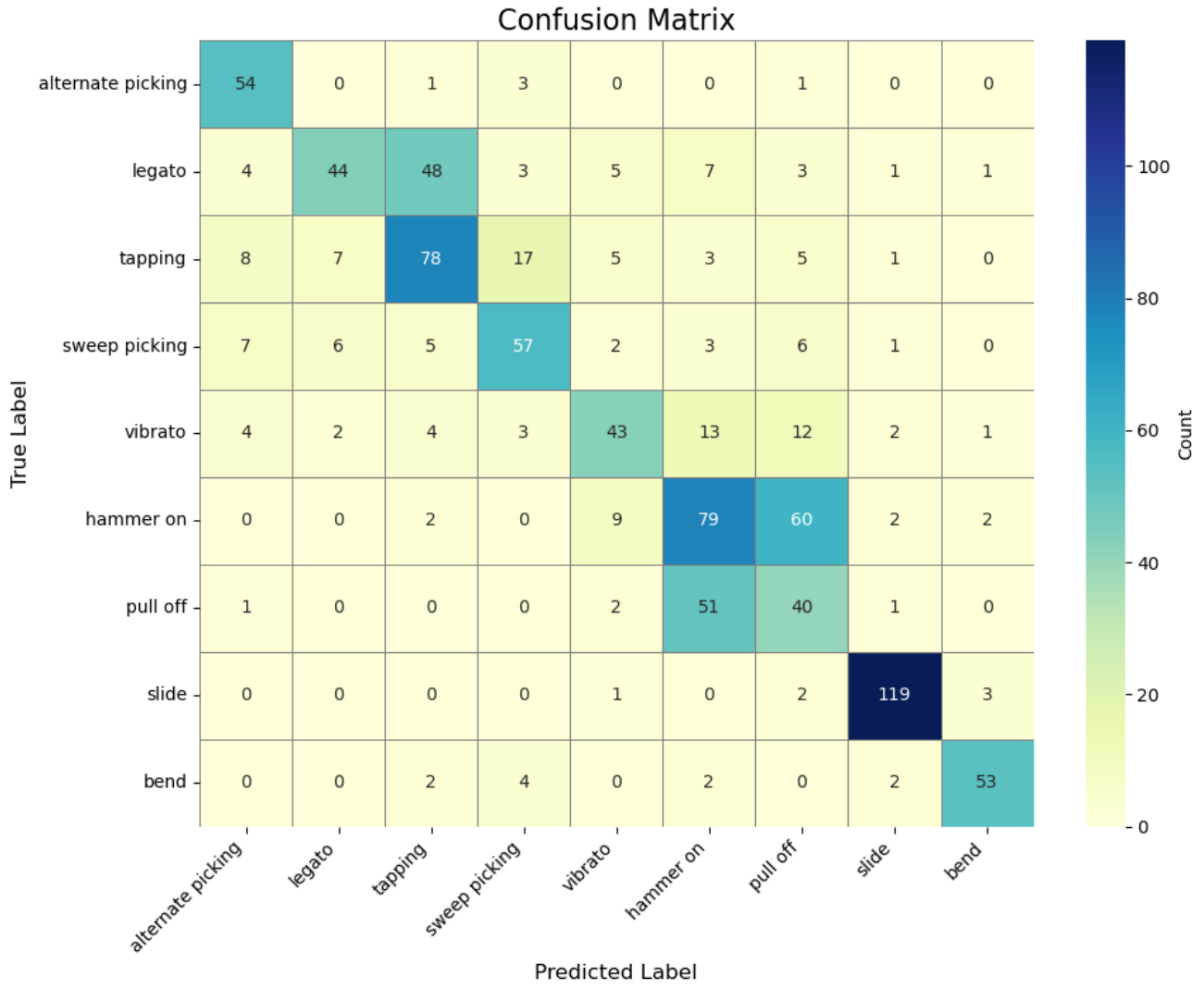
Still prove Wav2vec can adapt to different **sound** recognition/ classification tasks

Learning Framework: Wav2Vec 2.0 + linear classifier

Classification Report:

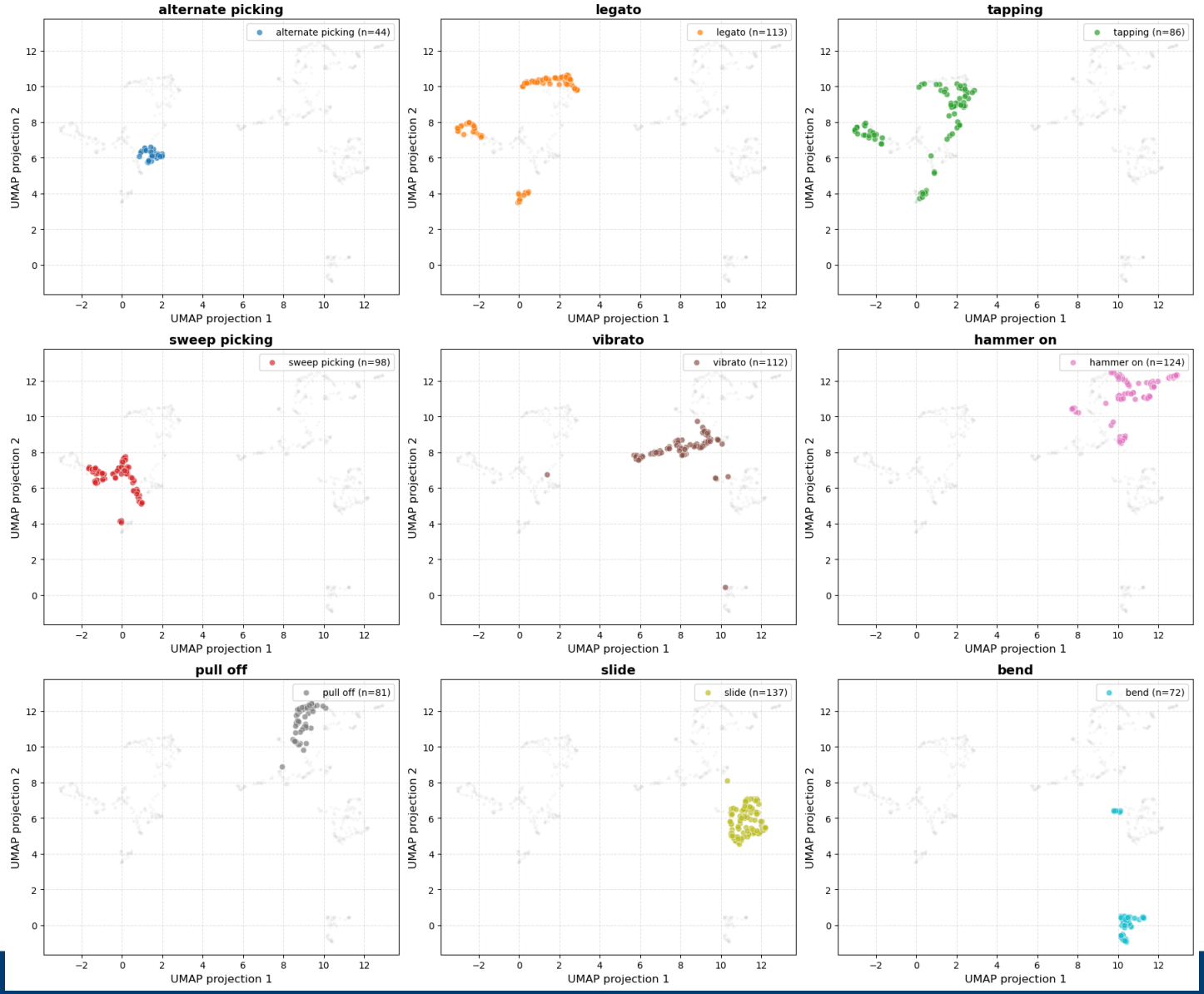
	precision	recall	f1-score	support
alternate picking	0.69	0.92	0.79	59
legato	0.75	0.38	0.50	116
tapping	0.56	0.63	0.59	124
sweep picking	0.66	0.66	0.66	87
vibrato	0.64	0.51	0.57	84
hammer on	0.50	0.51	0.51	154
pull off	0.31	0.42	0.36	95
slide	0.92	0.95	0.94	125
bend	0.88	0.84	0.86	63
accuracy			0.63	907
macro avg	0.66	0.65	0.64	907
weighted avg	0.64	0.63	0.62	907

- Misclassification more. Decreased on general performance but prove how powerful the wav2vec is.
- No guitar sound was pre-trained in Wav2vec, but it prove wav2vec can be used in different sound classification task.



Visualization on feature space (ConvNetwork base)

- Each class form groups before classification, decision boundaries are clear.
- Some overlaps between some classes (e.g., 2. legato & 3. tapping).
- Note that these features are compressed from higher dimensional space.
- A non-linear boundary could be found in higher dimensional laten space.



Plans for the Next Step

- Try some other combinations, play with different hyper-parameters inside the network to see if the overall accuracy can increase.
- Try another large pre-trained sound model, e.g., HTS-AT to evaluate if different feature embeddings can affect result.
- Record some other data (e.g., same technique but different guitar-amplifier setups played on our own) for testing, test the model's ability of generalization.

Thank you

References

[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.

How we divide our work

- *Ziyang Ou:* Literature review, code scripts, project report, slides, presentation.
- *Xiaohe Tian:* Code debugging, project report, slides, presentation.
- *Jiayin Yuan:* Project report, slides, presentation.

Data Augmentation Method

Original Dataset

• ~360 audio recordings in total

• Each recording is ~25 seconds and is already labeled with a **single technique**

→ *Multi-label classification is **not** considered for this study*

Offline augmentation: Pitch Shifting

• Randomly increase or decrease pitch

• Simulates wide pitch range of electric guitar

• Enlarged each class from ~40 → **~100 recordings**

• Each recording remains 25 seconds

• 5 sec trunk will then feed into training

Final sample size in total:

$$9\ classes \times 100\ clips \times \frac{5sec}{25sec} = 4500\ samples$$

On-the-fly (during training process) Augmentation:

Gain Alter

• Randomly increase or decrease audio gain

Goals of Augmentation Strategy

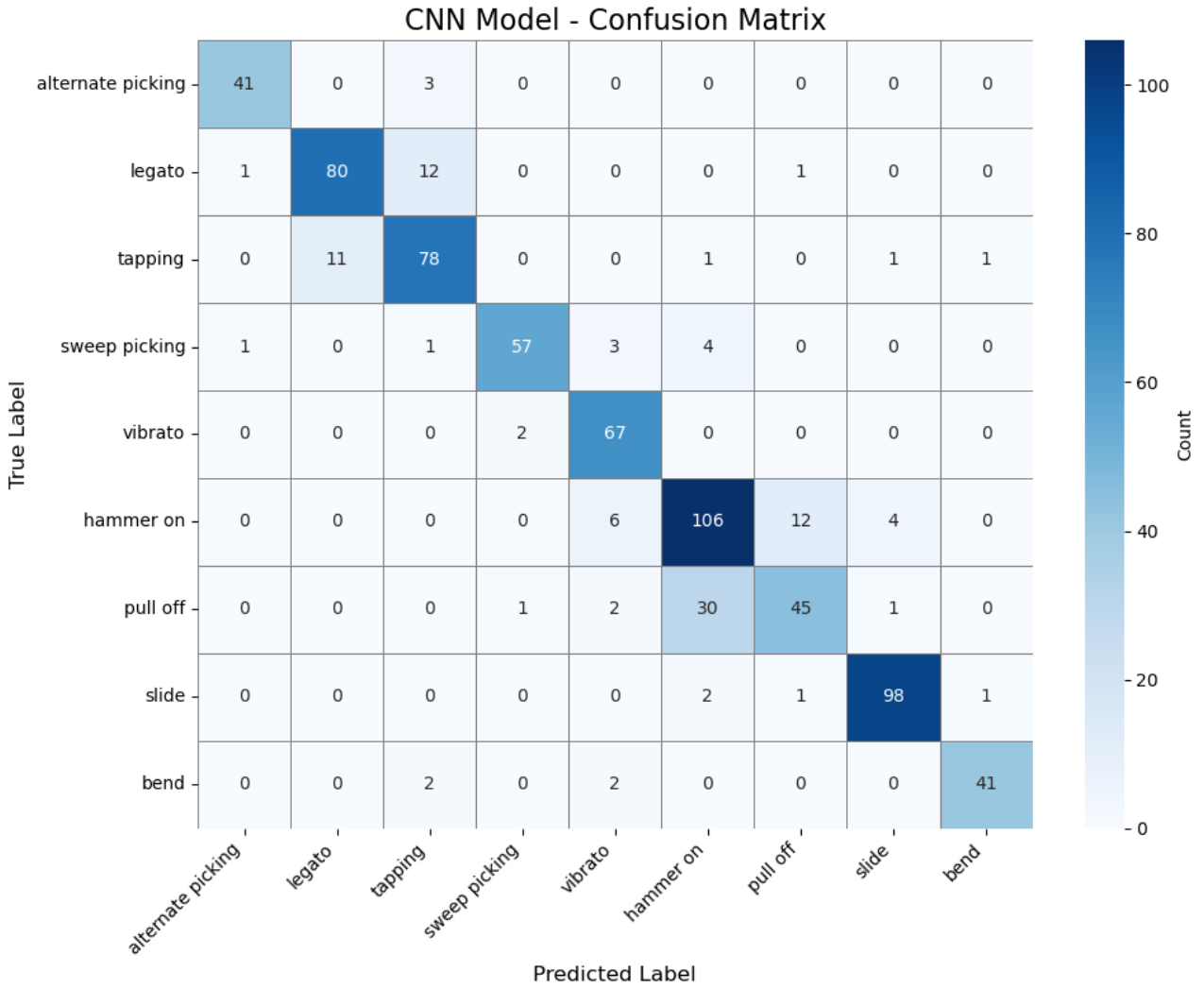
• **Broaden pitch coverage**, reflecting real-world recording situation

• **Increase sample quality and quantity**

• **Improve model generalization** and reduce overfitting

Performance evaluation

From testing, we get the confusion matrix showing numbers of predicted output labels vs. ground truth labels.



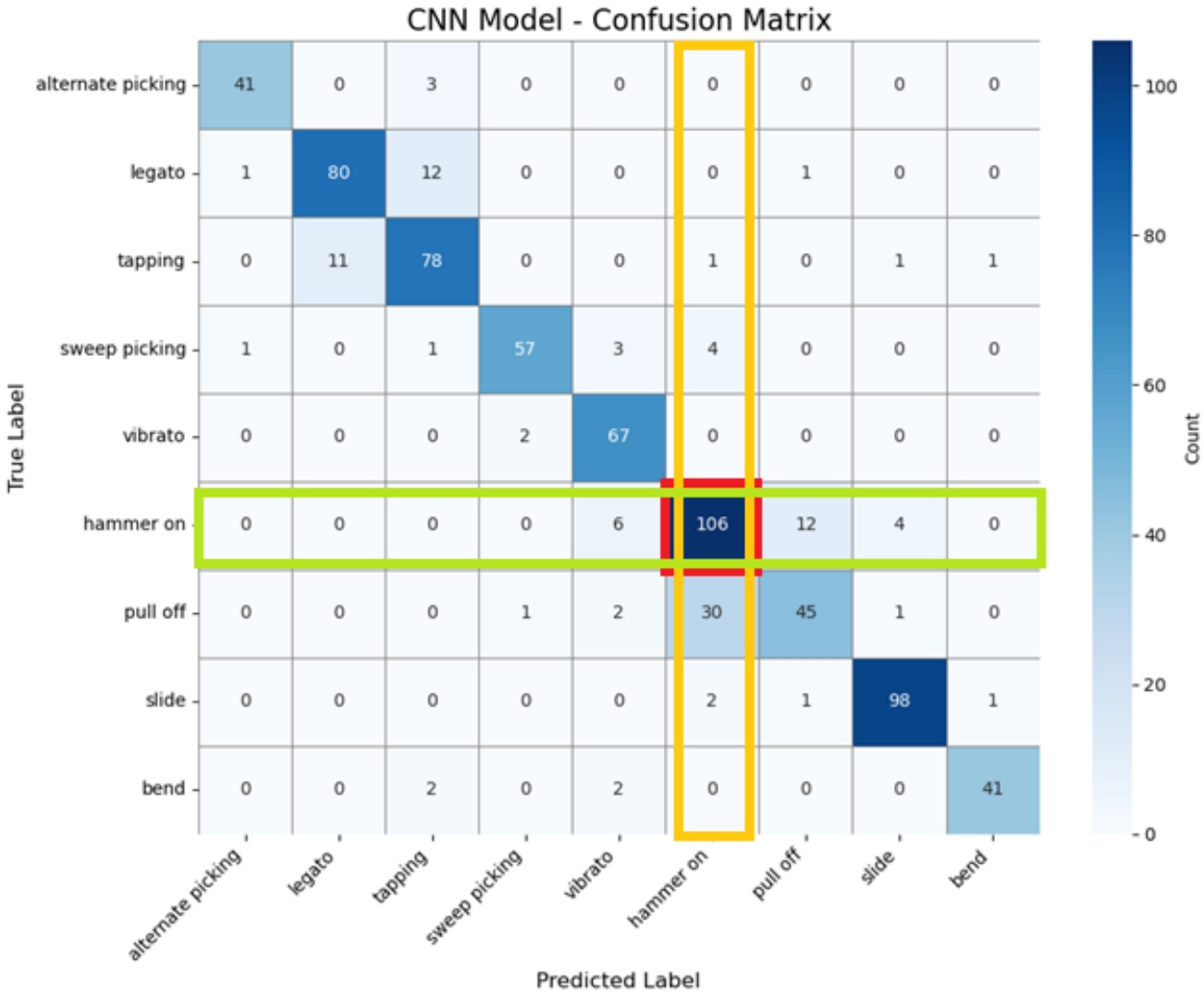
Performance evaluation

Performance metrics:

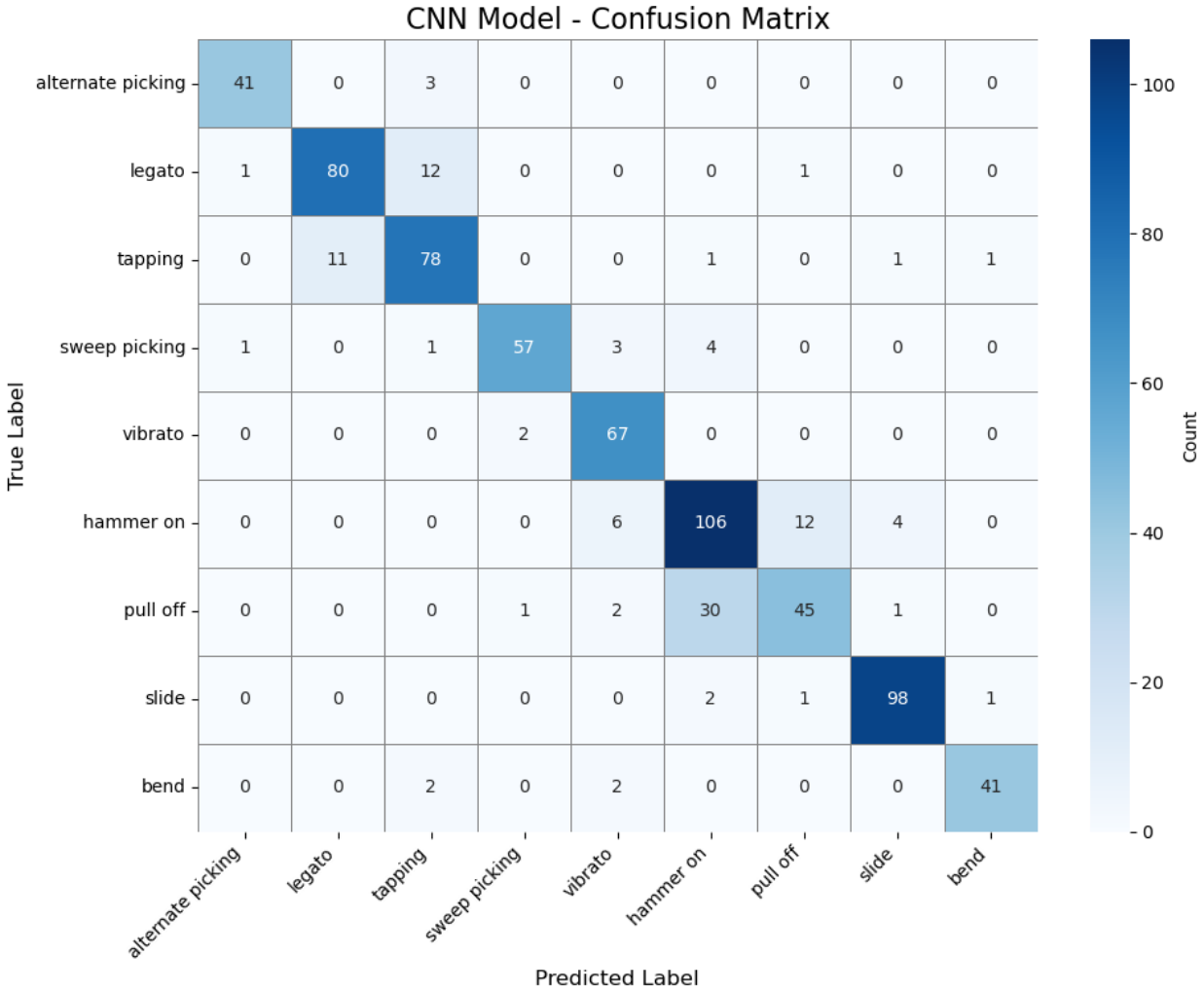
$$\text{Precision} = \frac{N(\text{correctly classified})}{N(\text{total predicted})}$$

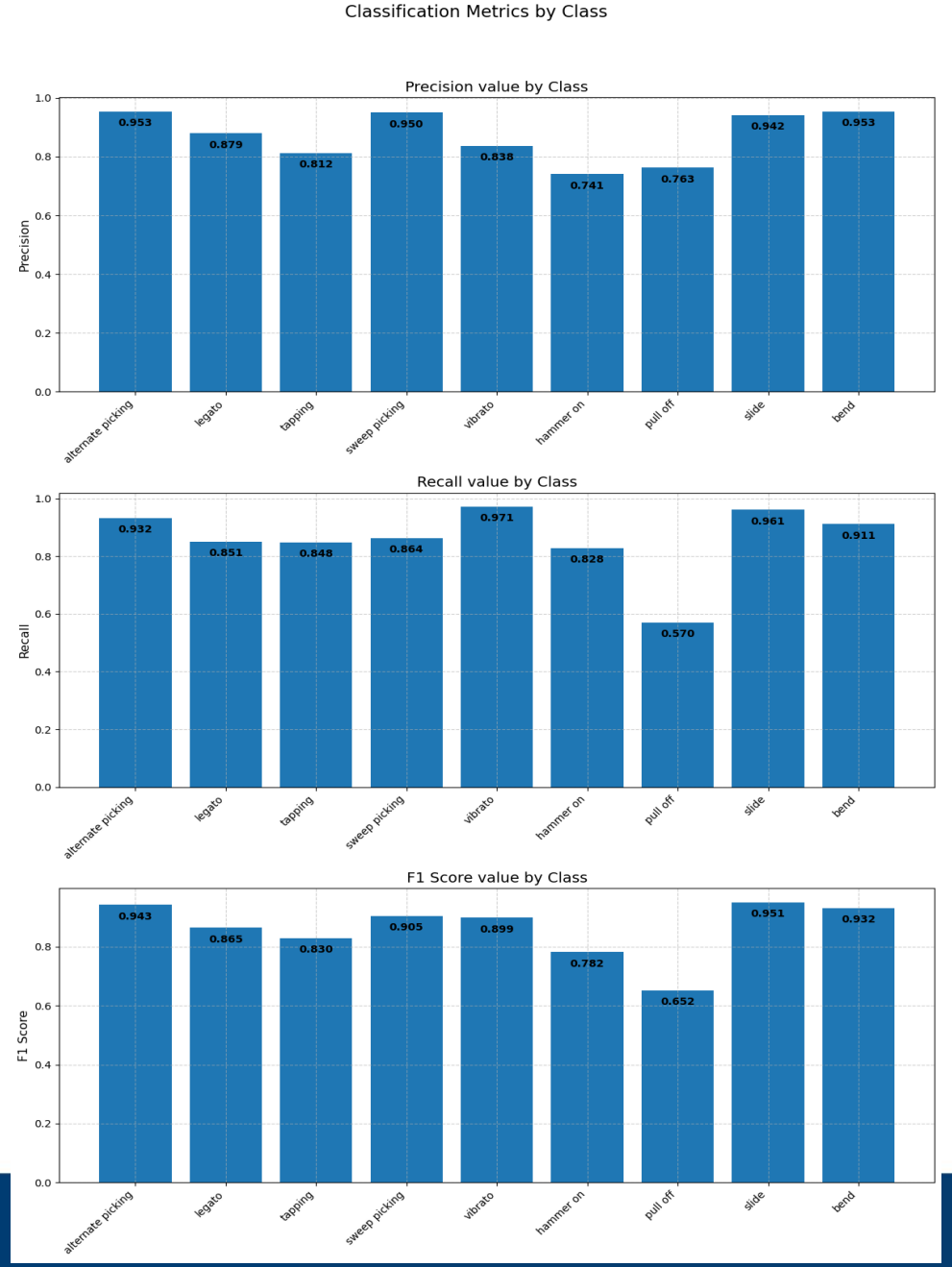
$$\text{Recall} = \frac{N(\text{correctly classified})}{N(\text{total ground truth})}$$

$$\text{F1-score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$



Performance evaluation



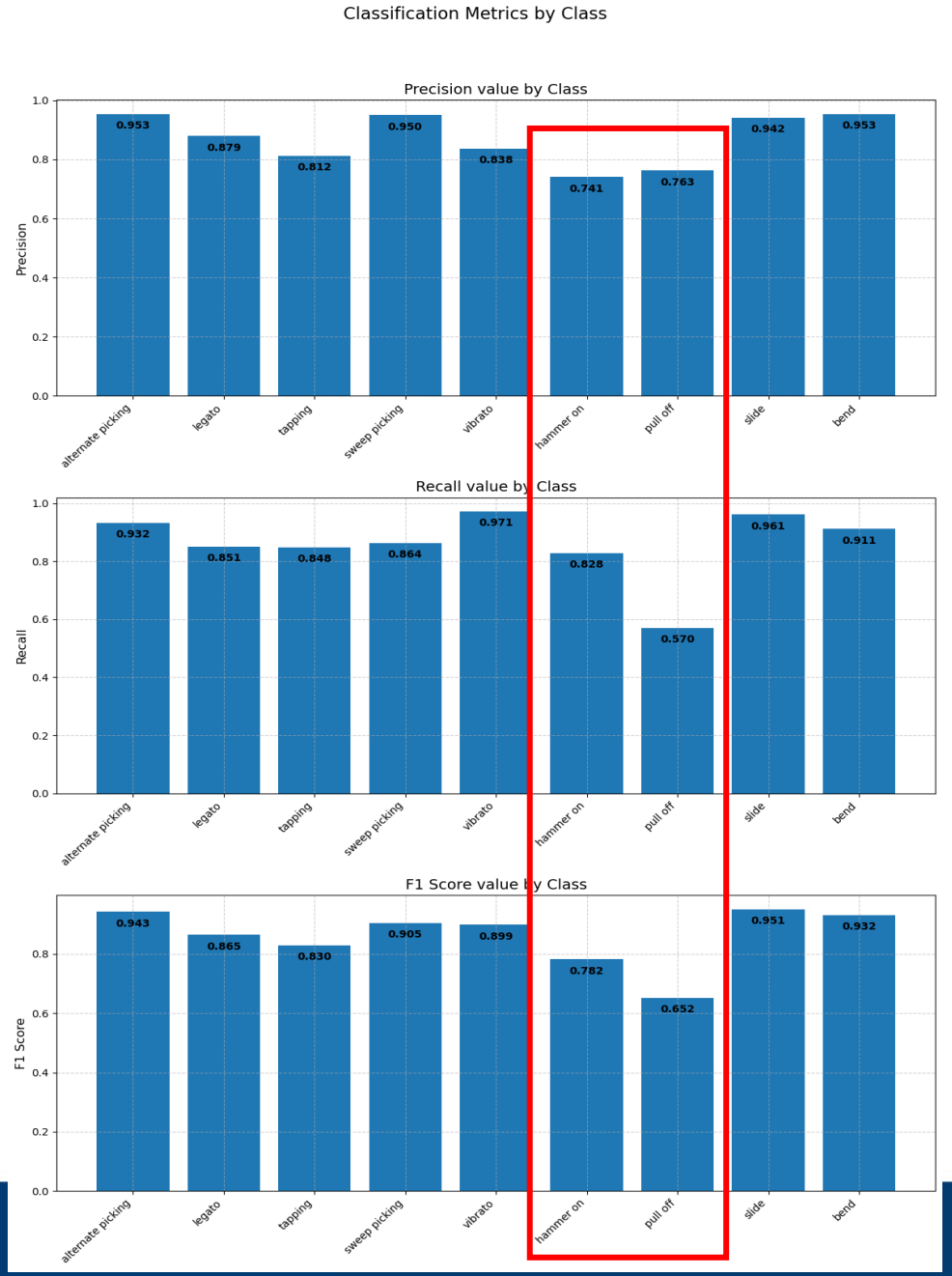


Performance evaluation

For Mel Spectrogram model:

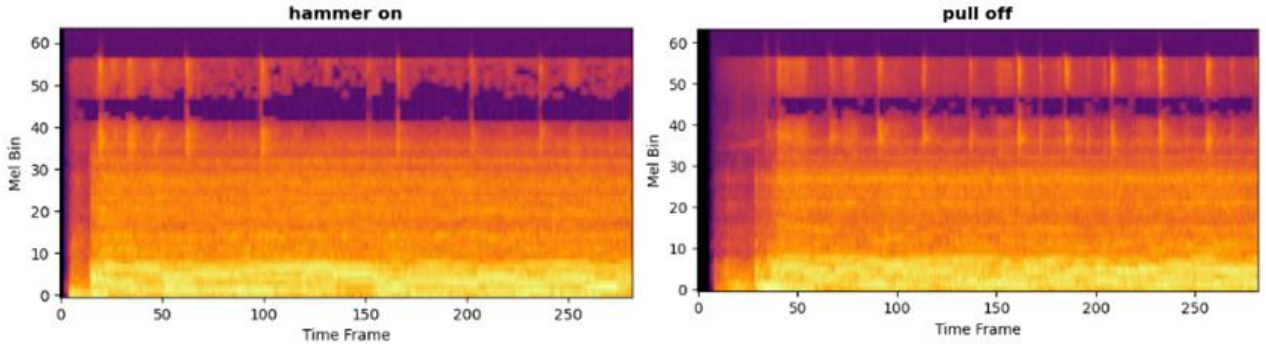
Excellent or good performance for most labels

Lower performance for “hammer on” and “pull off” due to mixing up of classifications of these two labels

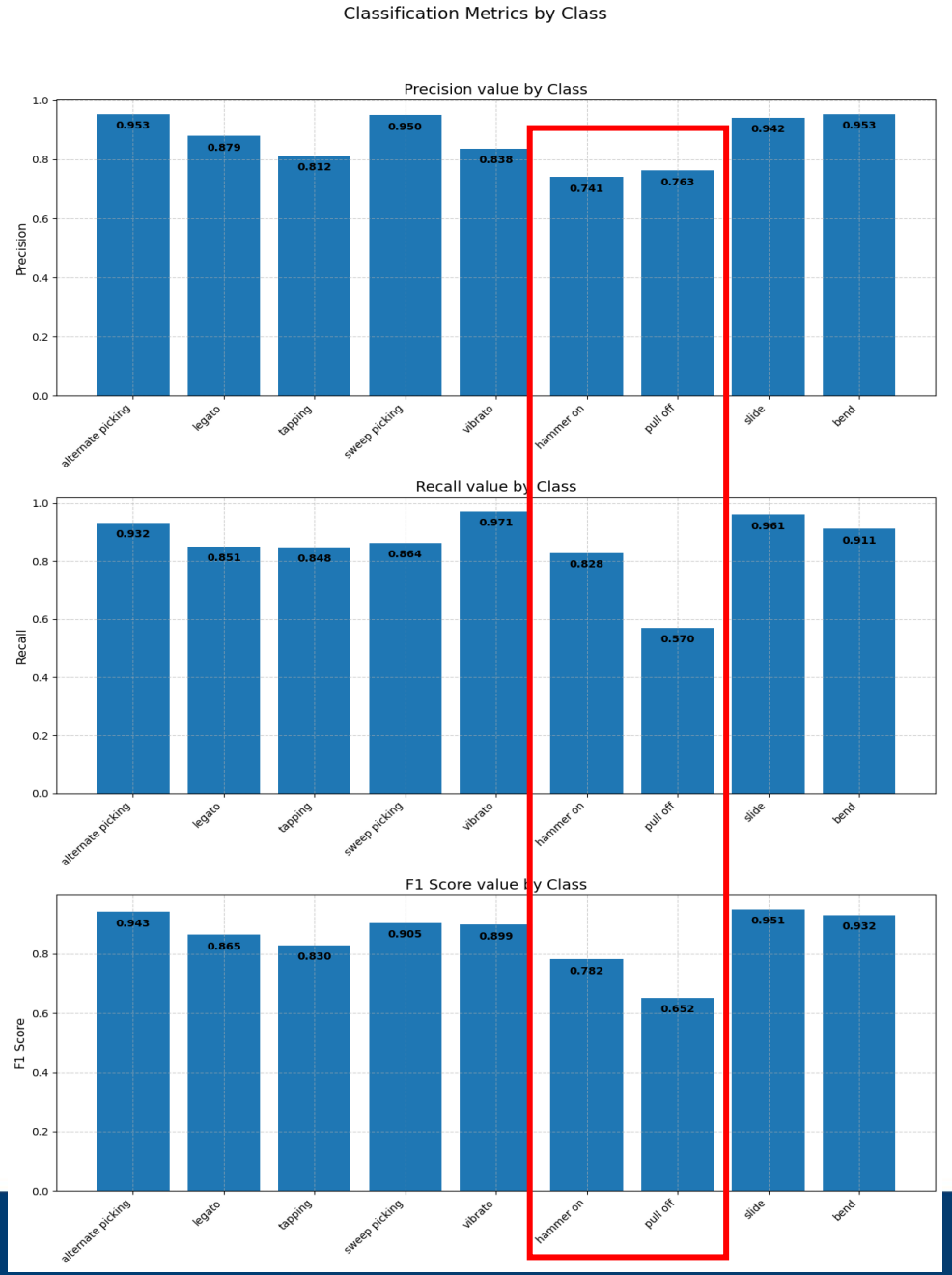


Performance evaluation

From Mel Spectrogram patterns:



We still can tell the subtle difference between them by ourselves, so there is room for improvement



Performance evaluation

For wav2vec model:

Generally, worse performance than Mel Spectrogram model

Therefore, we have verified the feasibility of not only machine learning, but also Mel Spectrogram