Classification of Instrument Timbres in Classical Music Sound Samples Using a Conformer Model Architecture

BROWN
Computer Science

CSCI 1470 FINAL PROJECT (TEAM LAZS): SARAH BAO, ZIQI FANG, ANDREW GAO, LUKE ZHAO

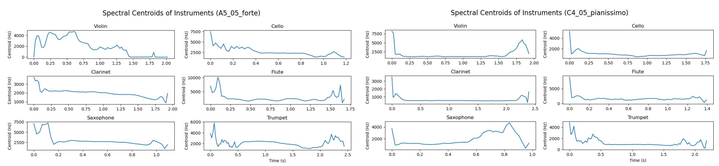
Introduction

We implemented a deep learning model using conformer (convolution + transformer) blocks which contained convolution and multi-headed attention layers to classify various instrument timbres. Our goal was to find meaningful relationships between timbres of instruments of different families.

Methodology

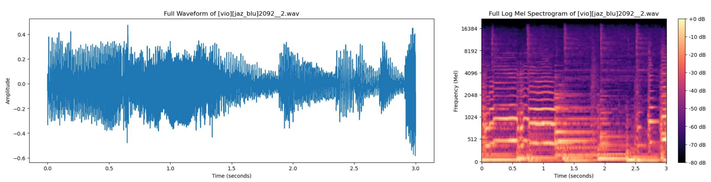
DATASET

The two datasets that we trained our model on are the London Philharmonic Orchestra dataset and the IRMAS dataset. The LPO dataset is a professionally recorded audio samples curated from the archives of the LPO with 13,681 mp3 files labeled across 20 instruments grouped into 4 categories. The IRMAS dataset consists of 9,579 wav files sampled at 44.1 kHz labeled with the predominant instrument (11 in total).



PREPROCESSING

All audio files were downsampled to 16 kHz, converted to mono, and segmented into 3-second clips. Log-mel spectrograms with 80 frequency bins and 300 time frames were computed for each clip, providing a compact but expressive time-frequency representation.

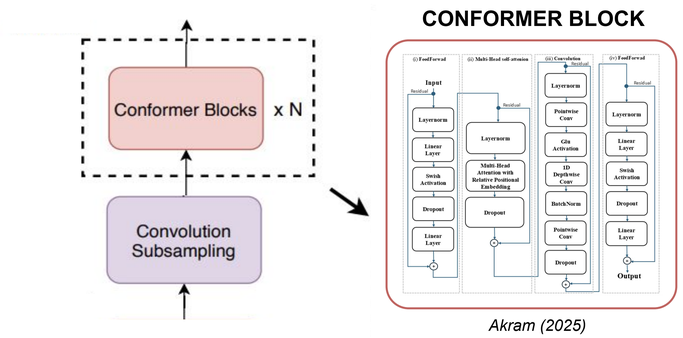


TRAINING

We trained our Conformer-based model using the Adam optimizer with an initial learning rate of 1e-3 and categorical cross-entropy loss with label smoothing of 0.1 to improve generalization. The training was conducted over 20 epochs with a batch size of 32, using 80-dimensional log-mel spectrograms extracted from audio recordings (LPO and IRMAS datasets). The model was validated on a held-out 20% subset of the data.

Model Architecture

Our model is based on the Conformer, a hybrid design that combines convolutional layers with self-attention to capture both local and global dependencies in the audio signal. Each input spectrogram is first linearly projected to a 256-dimensional feature space, then passed through a stack of four Conformer blocks, each comprising feedforward modules, multi-head attention with relative positional encoding, and convolutional sublayers with gated linear units. Below is a visualization of a singular conformer block:

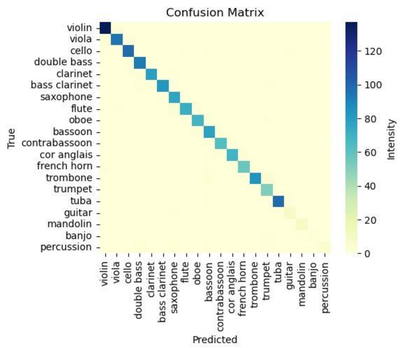


Akram (2025)

The output is aggregated over the time dimension and passed through a final dense layer followed by a log-softmax activation, producing class probabilities across all instrument categories.

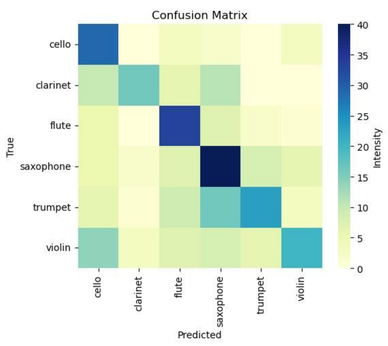
Results

LPO DATASET



Precision (P)	0.960
Recall (R)	0.917
F-1 Score	0.926
Accuracy	0.967

IRMAS DATASET



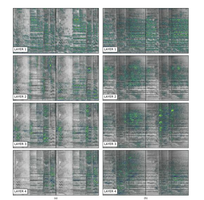
Precision (P)	0.555
Recall (R)	0.529
F-1 Score	0.511
Accuracy	0.516

Conclusion & Discussion

- With the Conformer Model training and testing on the LPO dataset, we achieved overall Precision (P) of 0.960, Recall (R) of 0.917, F-1 Score of 0.926, and Accuracy of 0.967, indicating that our model has outperformed the models in the original paper (Hernandez-Olivan et al.). This lead us to train and test the model on a more difficult data set (IRMAS dataset, whose sound files contain multiple instruments), we achieved overall Precision (P) of 0.555, Recall (R) of 0.529, F-1 Score of 0.511, and Accuracy of 0.516, which met our base goal.
- The use of conformer modules that combined convolution and attention addressed the problems brought up in the original paper that only implemented MLPs and FC models with Self-Attention components.
- Analysis of the Confusion Matrix allowed us to confirm a phenomenon that previous authors have discovered - instruments are more likely to be misclassified as other instruments in the same family/class. For instance, the violin/cello and clarinet/saxophone pairings are the most obvious from the confusion matrices.

Challenges & Future Work

- The IRMAS dataset consisted of sound files with multiple instruments but labeled with a single instrument. The task of predicting a predominant instrument in this situation is a task that is even challenging to a classically trained musician, which explains why we could not achieve our target and stretch goals after days of extensive tuning.
- The LPO dataset is also labeled with dynamics (on a spectrum from fortissimo to pianissimo) and bowing techniques (normal, arco, sul tasto). We could expand our model to classify the data with these labels.
- In implementing the Multi-Headed Attention Layers, we could extract the attention weights, and overlay them with the Mel-Spectrograms in order to visualize parts of the input data that the attention layers focused on. We could also visualize the weights in the convolution layers, all in hopes of developing a better understanding of why our model performs the way that it does.



Han, Kim, Lee (2017)

References

Akram, Muhammad Waseem, et al. "ChordFormer: A Conformer-Based Architecture for Large-Vocabulary Audio Chord Recognition." arXiv preprint arXiv:2502.11840 (2025).

Bosch, Juan J., Jordi Janer, Emilia Gómez, and Perfecto Herrera. "A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals." Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR), 2012.

Hernandez-Olivan, Carlos, and Jose R. Beltran. "Timbre classification of musical instruments with a deep learning multi-head attention-based model." arXiv preprint arXiv:2107.06231 (2021).

Y. Han, J. Kim and K. Lee, "Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 1, pp. 208-221, Jan. 2017