STRATEGIES FOR MUSICAL INSTRUMENT CLASSIFICATION

Grant Phillips

College of Natural Sciences
The University of Texas at Austin

ABSTRACT

Music Information Retrieval is a branch of Machine Learning that covers the analysis and generation of music. This report explores different strategies used to classify musical instruments from audio recordings. Classifying musical instruments can be used for several tasks, such as automated tagging of recordings for archiving and assisting with post-production editing of music. Some techniques explored were a Convolutional Neural Network and also a Transformer architecture where both accepted the Mel spectrogram representation of the audio signal. Acoustical and musical terminology is also covered to provide insight into the features of the instruments and how the models can extract them. The results of the experiments are provided to show how well each model performed.

1. INTRODUCTION

In this paper, two different machine learning models are explored for classifying musical instruments from recordings. Both models use a visual representation of audio recordings, called a Mel Spectrogram. A spectrogram is a way to view audio signals not only on the time domain but also the frequency domain. The horizontal domain can be thought of as the time domain while the vertical domain can be thought of as the frequency domain. The spectrogram is created by using the Fast Fourier Transform, a process of breaking down a signal into its sinusoidal components of varying frequencies by viewing the signal in discrete windows ("FFT", n.d.). A Mel Spectrogram is a variant of a spectrogram where the frequency domain is scaled to create a more linear representation. For illustrative purposes, the difference for human perception of 200 hZ to 400 hZ is much more noticeable than 15,000 hZ to 15,200 hZ. By applying the Mel scale to the spectrogram, it becomes more representative of human perception of frequency (Roberts, 2022). Converting the audio signal data into a visual representation allows for features to be extracted with similar methods as extracting features in images. The two models that were used

to extract features from the spectrograms were a convolutional neural network and a transformer network. Both models performed well in classifying solo instruments.

2. RESEARCH BACKGROUND

To classify musical instruments from audio recordings, it is important to make meaning from the audio signal to extract features. Since audio signals are only a temporal representation of the sound event, a spectrogram can be used to extract the spectral features of the sound event. Although the spectrogram is a visual representation of audio, patterns and features can still be extracted (Yu et. al, 2008). When using convolutional neural networks, the features of the spectrogram are more localized. Different methods can be used to extract features over the full frequency spectrum at given times (Lu et. al, 2021). The transformer can further develop meaning beyond localized features and instead create relationships between features inside and outside of local events. However, the convolutional neural network used in this research did perform close to the transformer model used. In the next few sections, different background terminology and concepts will be covered before reviewing the experiments and results.

2.1 ACOUSTICAL TERMINOLOGY

Before covering the models used in this research, it is important to cover acoustical terminology to make better sense of what the models are looking for in features and how they use those features to classify which instrument is playing in a recording.

To begin, **timbre** is the unique characteristic sound that an instrument makes. It is the reason why playing one note on one instrument sounds different when playing the same note on another ("What is Timbre in Music?", n.d.). **Color** is synonymous with timbre when describing how an instrument sounds.

Frequency, measured in hertz (Hz), is the speed of the oscillating waveform. A higher frequency produces a higher pitch than a lower frequency. For example, playing the lowest key on a piano has a lower frequency than playing the highest key. Based on the Fourier Theory, a signal can be broken down into a summation of different sinusoid signals at varying frequencies

(Lostanlen et. al, 2019). This means that the timbre of an instrument is a combination of different simultaneous sinusoid frequencies playing which give the instrument its characteristic sound.

Harmonic sound is when the partial frequencies of a base frequency are whole-number multiples of the base frequency, while **Inharmonic** sound is when the partial frequencies do not have a similar relationship (Pellman, 1994). A harmonic sound could be a trumpet or piano, where the note can be discerned. An inharmonic sound could be a snare drum, bell, clap, or cymbal crash. An example of a harmonic sound is one that has a base frequency of 200 Hz, and partials 400 Hz, and 800 Hz. An example of an inharmonic sound is 200 Hz, 213 Hz, 332 Hz, and so on.

Acoustical Impedance is the ratio of pressure to volume flow (Dickens et. al, 2007). It is used to measure how an instrument will respond to different frequencies of air pressure ("What is acoustic impedance?", n.d.). It is particularly useful for wind instruments that have a column of oscillating air. The way the instrument responds to these different oscillating frequencies is one aspect of the timbre.

2.2 INSTRUMENT CHARACTERISTICS

Each instrument has a different set of characteristics that create its timbre. The clarinet for instance has many elements that build its timbre, such as lip pressure, embouchure (the shape of the mouth on the reed), and register, which is a set of tonal ranges in the instrument (Almeida et. al, 2013). Figure 1 is a random sample of a clarinet from the Medley Solos DB converted to a Mel spectrogram.

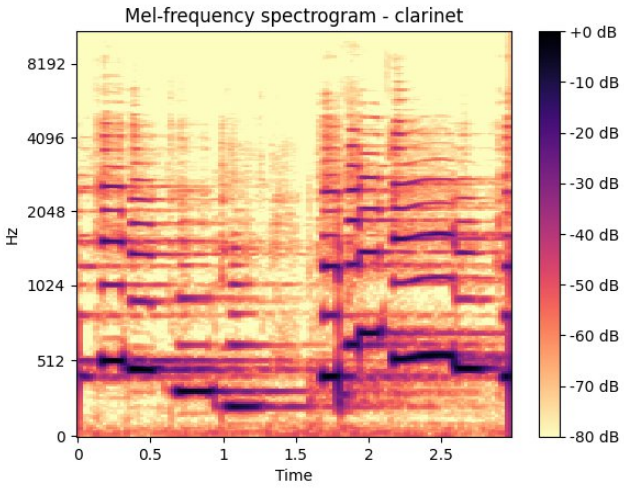


Figure 1. Mel Spectrogram of a clarinet

The image's x-axis is the time span of three seconds. The y-axis is the frequency domain in the Mel scale. The fundamental and partial frequencies are discernible, noted by the dark purple colors. The darker the color, the stronger the amplitude of the frequency is.

Another random sample, this time of a piano, can be seen in Figure 2. A clarinet can only produce one note at the same time, known as monophonic, while the piano can play multiple notes simultaneously, known as polyphonic. Since the piano can play more than one note simultaneously, it is possible that the Mel spectrogram shows several frequencies playing simultaneously, which can be seen in this case. It can also be assumed that the performer was playing blocked chords, where each key is pressed at the same time, due to the attack designated by the vertical lines of where the frequencies start. At times 1, 1.5, and just after 2 seconds, this can be seen. Another feature of the piano is the dying sound of a key being held. The strings of a piano are hammered at each keystroke so the sound naturally rolls off if the key or sostenuto pedal is held. After 2 seconds at the top of the frequency spectrum, the higher frequencies die off faster than the lower frequencies.

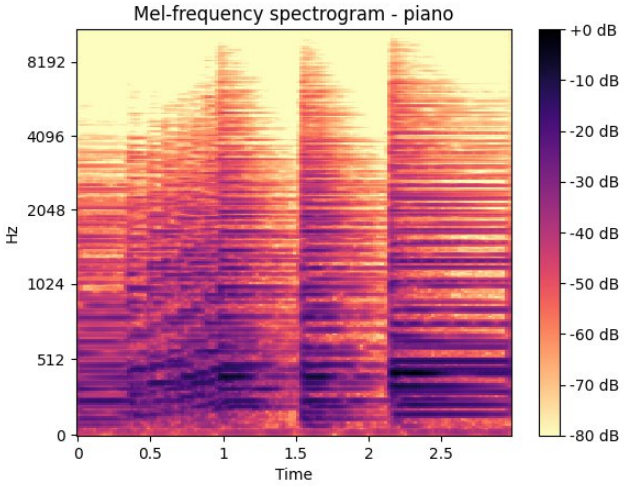


Figure 2. Mel Spectrogram of a piano

The electric distorted guitar has a dissimilar spectrogram from the clarinet and piano. Although it has defined harmonic frequencies, there is also a lot of noise as seen by the darker background of Figure 3. This can be thought of as inharmonic features.

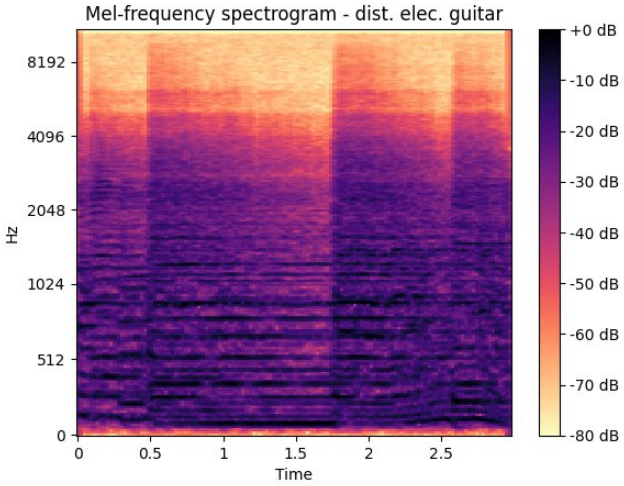


Figure 3. Mel Spectrogram of a distorted electric guitar

When comparing the spectrogram images, the flute has similar features as the clarinet. As will be seen later, both transformer and convolutional neural network models struggled with classifying between the two. A different study performed by Antti Eronen for classifying musical instruments also had struggling predictions between the flute and Bb clarinet (Eronen, 2001). This could be due to the similarity of the features. However, the instruments do have different characteristics. For one, both instruments have a different pitch range. A flute has a range of B4 to D7 while the Bb clarinet has a sounding range of D3 to G6, i.e. the flute has a higher range than the clarinet (Adler, 2002). Also, as previously mentioned, the clarinet has a different set of ranges, called “registers” that have their own characteristic sound (Almeida et. al, 2013).

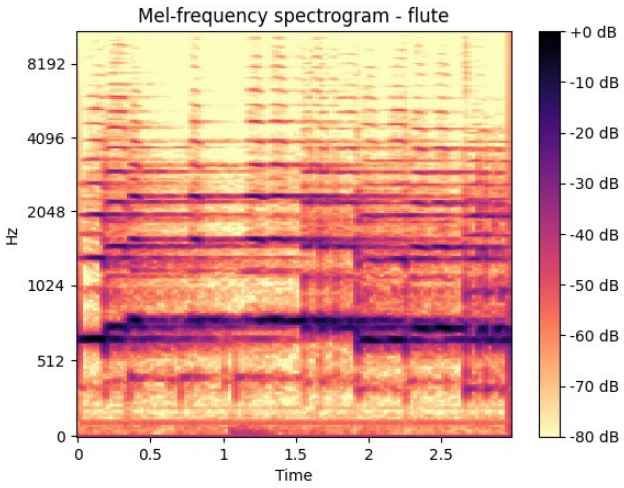


Figure 4. Mel Spectrogram of a flute

The tenor saxophone, like the flute and clarinet, is part of the woodwind family of instruments. It is similar in tone to the clarinet, which is also played with a reed in a mouthpiece (Adler, 2002). One key difference between the clarinet and saxophone is the order of harmonics. The clarinet has a cylindrical bore while the saxophone has a conical bore. The cylindrical bore of the clarinet produces maximal acoustical impedance at its odd harmonics, while the saxophone produces maximal impedance at both even and odd harmonics (Pipes, 2018). Figure 5 shows a Mel spectrogram sample of a tenor saxophone performance.

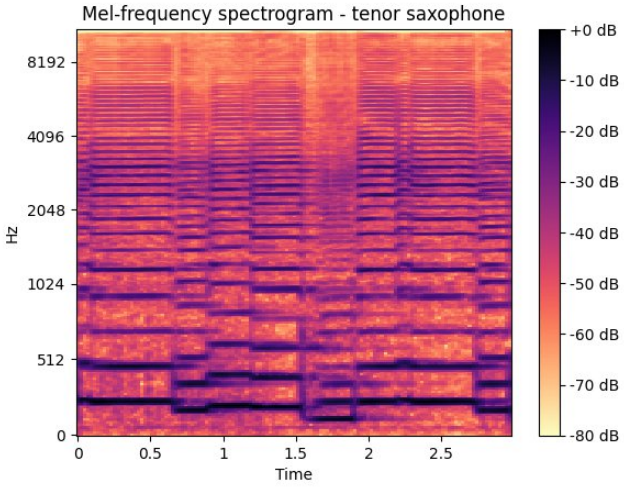


Figure 5. Mel Spectrogram of a tenor saxophone

Figure 6 shows a three second sample of a female singer. Some features of vocals are that it is a monophonic instrument, formant frequencies can be adjusted by mouth shape to create vowels, and the larynx can be adjusted to have an affect on the formant frequencies (Sundberg & Nordström, 1976). In addition, inharmonic sounds can be created, such as breath sounds and clicks. The diverse range of effects the voice has can create many features that can be extracted.

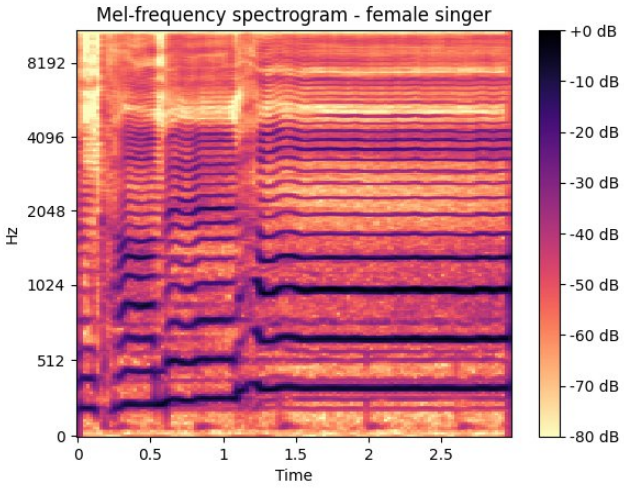


Figure 6. Mel Spectrogram of a female singer

Brass instruments such as the trumpet are considered a closed-pipe instrument, where the lips are the closed-end. The effect of a closed pipe is that the odd harmonics are present instead of all harmonics (Hartmann, 2013). However, with the bell of the instrument, all harmonics are present. Also, the lowest frequency is low in amplitude compared to the other harmonics (Hartmann, 2013). Figure 7 shows how all harmonic series are present, not only odd-numbered harmonics.

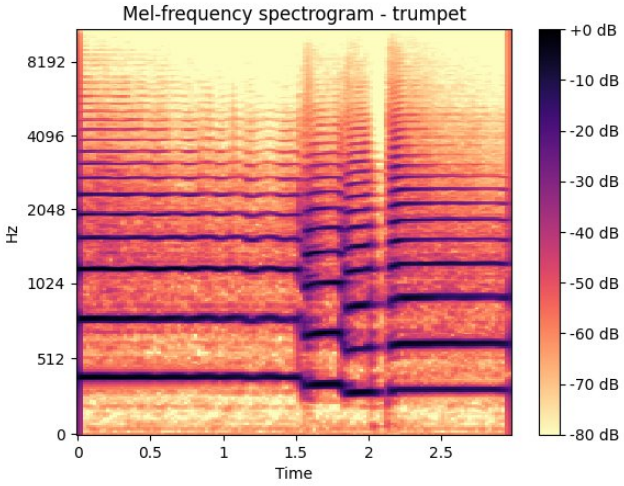


Figure 7. Mel Spectrogram of a trumpet

The violin, although typically performed as a monophonic instrument, can play multiple notes at once. This technique is called “double-stop” when two strings are simultaneously played and “triple-stop” when three are simultaneously played. Also, since it is a bowed instrument, the dynamics can be controlled. Compare this to another instrument such as the piano which does not have the same level of dynamic control. In Figure 8, this can be seen in a passage where the violinist is running through multiple notes.

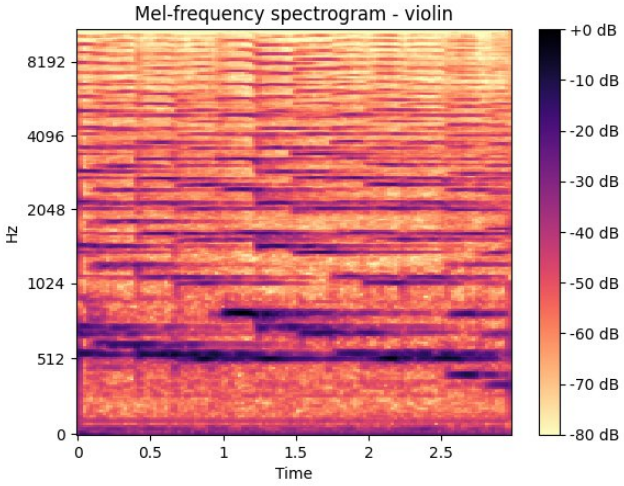


Figure 8. Mel Spectrogram of a violin

2.3 AUDIO CLASSIFICATION MODELS

The first model explored in this research was the convolutional neural network. With a convolutional neural network, a kernel is superimposed with repeated convolution operations to extract features in the data (Zhang, 2021). These type of networks are typically used for image and classification tasks ("What are Convolutional Neural Networks?", n.d.). The goal with using a convolutional neural network in this research is to see if features can be extracted when treating the audio signal as an image via the Mel spectrogram. Similar techniques for feature extraction via a convolutional neural network were performed by Jingwen Zhang. Unlike the research performed in this paper, Zhang's research used a one-dimensional convolution instead of a two-dimensional convolution. They found that by using a one-dimensional version, which focuses on the time dimension, the convolution had a closer relationship with the audio signal, which is itself temporal based (Zhang, 2021). However, with this research two-dimensional convolutional layers were used since the audio signals were being treated as images with the x-axis representing time and the y-axis representing frequency.

The second model explored in this research was the transformer. Since audio is temporal based, sequential modeling was sought after which the transformer architecture provides. The goal of the model is to classify an input, so a transformer encoder was used without any decoding (Gong et. al, 2021). A transformer architecture specifically used for audio classification on spectrogram inputs was designed by Gong et. al and was used in this report. The specific model used was the **ASTForAudioClassification** model from Hugging Face ("Audio Spectrogram Transformer", n.d.). The spectrogram that is fed into this model is split into several overlapping patches before being projected onto an embedding layer. This embedding layer is added to a positional embedding layer before being fed into the transformer layer. The purpose of the positional embedding layer is to have the model make sense of the ordering of acoustical events (Phuong & Hutter, 2022). It is built with a combination of sine and cosine functions over a vector for each position (Phillips, 2023). Attention is used in the transformer which is used to find relationships in different parts of the input with a key and query configuration (Vaswani et. al, 2023). With the attention layer, the model learns about relationships between different patches in the spectrogram to learn features. If the patches fed into the attention layer are similar, they could be sharing the same acoustical event (Sperber et. al, 2018). The next sublayer in the transformer is the feedforward network, which is used to allow for various paths to be possible from the

attention sublayer (Alammar, 2018). Finally, the output of the transformer encoder is fed into a log softmax function to calculate the probabilities of each musical instrument.

One area that was not explored is a combination of both convolutional neural networks and transformers. Won et. al proposed such a configuration for a music tagging network (Won et. al, 2021). This structure uses a convolutional neural network to first learn local acoustical features. Its output is then fed into a transformer layer which is used to model the sequences. Their approach to training the model was a semi-supervised one, where the training included both labeled and unlabeled data. Since the results of the convolutional neural network and transformer network previously mentioned had successful results, this type of architecture was not explored.

3. DATA

To train the classifiers used in this experiment, the dataset Medley Solos DB was used ("Medley-solos-DB: A cross-collection dataset for musical instrument recognition", n.d.). There are 21,572 recordings in this dataset of musical instruments, which includes clarinet, distorted electric guitar, female singer, flute, piano, tenor saxophone, trumpet, and violin for a total of eight instruments to classify. For each recording, there is a single instrument performing. To load the audio data, mirdata was used which contains a collection of multiple datasets used in the Music Information Retrieval field (Bittner et. al, 2019). Although the dataset already has split training, testing, and validation sets, the training and testing splits were generated randomly by 70% and 30% respectively to take advantage of the splitting features of the mirdata software.

The total amount of recordings per instrument are:

Instrument	Total Recordings
Piano	6032
Violin	5971
Flute	3555
Distorted Electric Guitar	1854
Female Singer	1744
Clarinet	1311
Trumpet	627
Tenor Saxophone	477

Due to the uneven distribution of data, a method for scaling the weights was used for the loss function (Thilakarathne, 2021).

1 - (number of samples of the class / total number of samples)

By using this newer distribution, less emphasis is placed on updating classes with a large amount of recordings compared to classes with fewer recordings. This allows updates to be more uniform.

Included in the dataset is a .csv file consisting of a row for each recording. The columns include which subset the recording belongs to out of test, training, and validation, the name of the instrument in the recording, an ID for the instrument in a numerical representation, an ID for the song, and a UUID for the row. Each recording is a total of three seconds long with only a single instrument playing. Since only a single instrument is performing, it is easier to extract features instead of having a mixture of multiple features from multiple instruments. To augment the training data, only two seconds are used when predicting a recording by randomly picking a starting position within the full clip ("Audio Data Augmentations—Music Classification: Beyond Supervised Learning, Towards Real-world Applications", n.d.).

4. METHODS

For both model types, a learning rate of .0001 was used and the training sessions iterated over 15 epochs. Before feeding the audio signal into the network, a few transformations were performed. First, the signal was originally using a sample rate of 44,100. To reduce the amount of parameters needed, the signal was resampled to a sampling rate of 16,000. Next, to artificially add more training data, a random two second sample was extracted from the three second sample. Finally, a transformation into a Mel spectrogram was performed using **ASTFeatureExtractor** from Hugging Face ("Audio Spectrogram Transformer", n.d.). The remaining parameters used were the default values for the extractor. The convolutional neural network used three convolutional blocks with input/output layers being 1 to 32, 32 to 64, and 64 to 128 as used by Valerio Velardo in a demonstration of convolutional neural networks for audio classification (Velardo, 2021). A

final linear layer was used to create an output for the total classes of 8 before being sent to a Log Softmax layer for log probabilities.

The transformer network used 12 attention heads, and 2 hidden layers. All other settings used the default values of the **ASTForAudioClassification** and **ASTConfig** classes created by Hugging Face. The final layer was also a Log softmax function to calculate the log probabilities for each class.

5. RESULTS

For the convolutional neural network, the f1-score for accuracy of the test set was 93% overall. The best performing classification was the violin, with a precision of 97%, a recall of 97%, and an f1-score of 97%. The worst performing classification was the clarinet, with a precision of 98%, a recall of 67%, and an f1-score of 80%. Table 1 has the full results of the scores for the convolutional neural network's performance. Table 2 has the full results for the transformer network. When comparing the results of the two networks, the transformer had a better overall performance compared to the convolutional neural network. However, the convolutional neural network had much better performance with precision of the clarinet, and recall of the female singer and flute. The transformer had much better performance in the recall of the clarinet, precision of the female singer and flute, and recall of the tenor saxophone.

Table 1. Convolutional Neural Network results

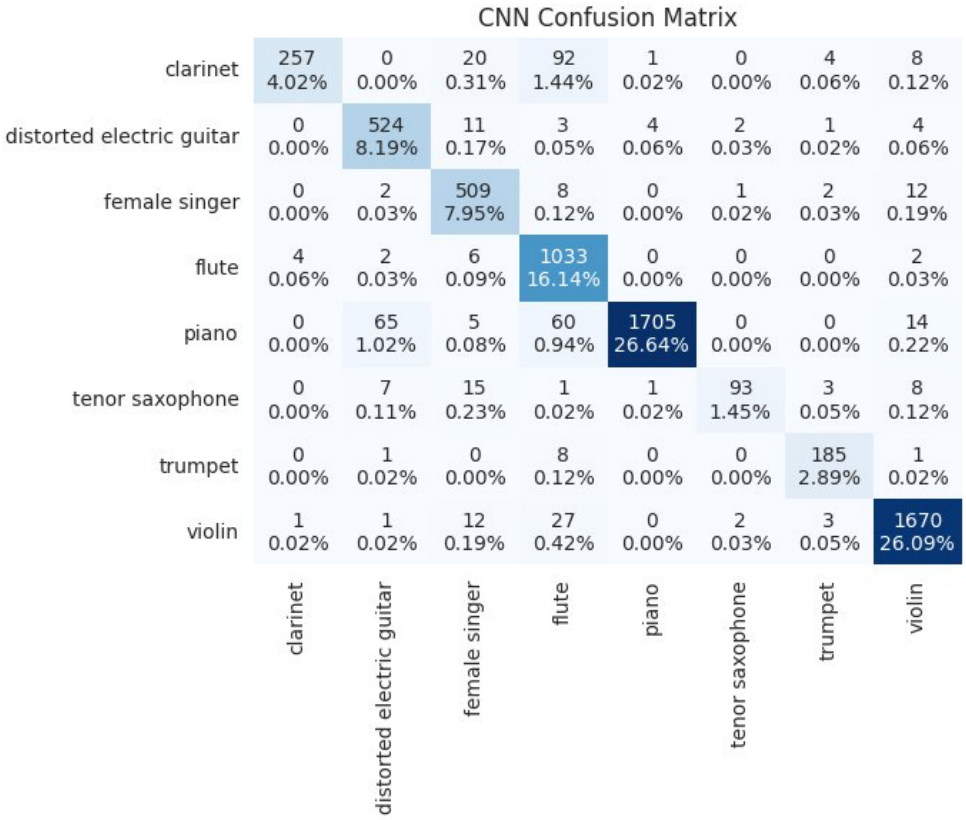
	Precision	Recall	F1-Score	Support
Clarinet	98%	67%	80%	382
Dist. Elec. Guitar	87%	95%	91%	549
Female Singer	88%	95%	92%	534
Flute	84%	99%	91%	1047
Piano	100%	92%	96%	1849
Tenor Saxophone	95%	73%	82%	128
Trumpet	93%	95%	94%	195
Violin	97%	97%	97%	1716
Accuracy			93%	6400
Macro average	93%	89%	90%	6400
Weighted average	94%	93%	93%	6400

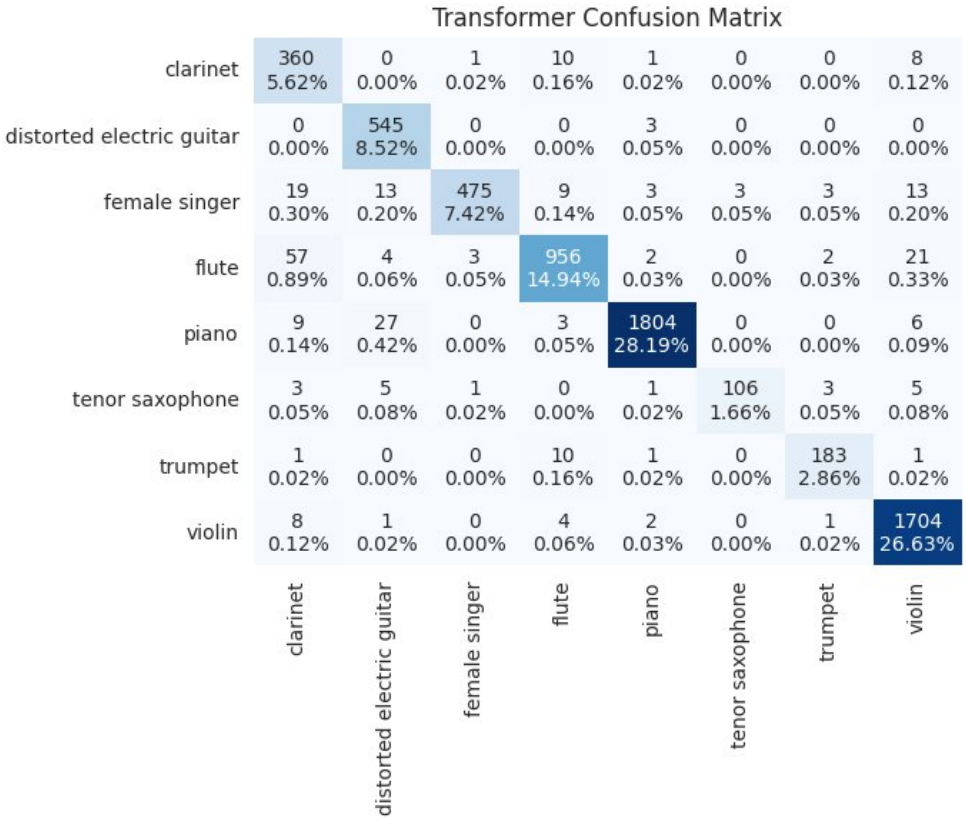
Table 2. Transformer results

	Precision	Recall	F1-Score	Support
Clarinet	79%	95%	86%	380
Dist. Elec. Guitar	92%	99%	95%	548
Female Singer	99%	88%	93%	538
Flute	96%	91%	94%	1045
Piano	99%	98%	98%	1849
Tenor Saxophone	97%	85%	91%	124
Trumpet	95%	93%	94%	196
Violin	97%	99%	98%	1720
Accuracy			96%	6400
Macro average	94%	94%	94%	6400
Weighted average	96%	96%	96%	6400

Table 3 shows the confusion matrix for both the convolutional neural network and transformer network. The columns represent the predictions for the given instrument and the rows represent the gold labels. One area of interest is the false positive amount for predicting the distorted electric guitar when the label was actually a piano. 1.02% of the cases in the convolutional neural network were this case while the transformer network only had 0.42% of the cases falling in this category. Figures 2 and 3 have some visual similarity due to the polyphonic nature of both instruments and the attack on each chord having a well defined vertical line. Another area of interest is the prediction of flute while the actual label was a clarinet. When comparing the convolutional neural network with the transformer network, 1.44% of the cases predicted flute when it was the clarinet while the transformer network only had 0.16% of the cases falling under this category. As previously mentioned, both instruments have tonal similarities in different registers of the clarinet. It is possible that the samples of the clarinet were in these registers so the model had difficulty differentiating between the two.

Table 3. Confusion Matrices





6. CONCLUSION

With this research, two different classification models were explored for solo musical instrument recordings. The transformer model performed slightly better than the convolutional neural network when trained under the same conditions and hyperparameters. This slight increase in performance could be due to the attention layer of the transformer, which builds relationships over time and frequency while the convolutional neural network builds relationships on a more local level.

There are many use-cases where this type of technology could be used. For example, a database of sound samples could use this to automatically tag recordings during uploading to remove the tedious process of listening to each recording and labeling manually. This can allow for a bulk import of recordings. Also, this type of technology could be useful in post-production. An editor could use this in post-production software to show exactly where an instrument starts and ends when working with an unedited recording of multiple instruments. For example, it could be useful while editing a recording of an orchestral performance; if there was a clarinet solo, the editor may want to find the location of that section of the recording quickly instead of manually listening to multiple sections. However, the model and training would have to be modified for this case. Since the models researched in this report are for solo instruments, a model that works with multiple instruments would have to be trained on a multi-instrument dataset such as IRMAS (Bosch et. al, 2014). The amount of instruments in the database would also have to be increased since there is only a finite amount of labels in both Medley solos DB and IRMAS.

To view the code used in this research, please visit:
https://github.com/GCPhillips/musical-instrument-classifier

REFERENCES

Adler, S. (2002). The Study of Orchestration (3rd edition). W. W. Norton & Company.

Alammar, J. (2018, June 27). The Illustrated Transformer. https://jalammar.github.io/illustrated-transformer/

Almeida, A., George, D., Smith, J., & Wolfe, J. (2013). The clarinet: How blowing pressure, lip force, lip position and reed "hardness" affect pitch, sound level, and spectrum. The Journal of the Acoustical Society of America, 134(3), 2247–2255. https://doi.org/10.1121/1.4816538

Audio Data Augmentations—Music Classification: Beyond Supervised Learning, Towards Real-world Applications. (n.d.). Retrieved November 16, 2023, from https://music-classification.github.io/tutorial/part3_supervised/data-augmentation.html

Audio Spectrogram Transformer. (n.d.). Retrieved November 12, 2023, from https://huggingface.co/docs/transformers/model_doc/audio-spectrogram-transformer

Bittner, R., Fuentes, M., Rubinstein, D., Jansson, A., Keunwoo Choi, & Kell, T. (2019). mirdata: Software for Reproducible Usage of Datasets. https://doi.org/10.5281/ZENODO.3527750

Bosch, J. J., Fuhrmann, F., & Herrera, P. (2014). IRMAS: a dataset for instrument recognition in musical audio signals (1.0) [dataset]. Zenodo. https://doi.org/10.5281/zenodo.1290750

Dickens, P., Smith, J., & Wolfe, J. (2007). Improved precision in measurements of acoustic impedance spectra using resonance-free calibration loads and controlled error distribution. The Journal of the Acoustical Society of America, 121(3), 1471–1481. https://doi.org/10.1121/1.2434764

Eronen, A. (2001). Comparison of features for musical instrument recognition. Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575), 19–22. https://doi.org/10.1109/ASPAA.2001.969532

FFT. (n.d.). Retrieved November 16, 2023, from https://www.nti-audio.com/en/support/know-how/fast-fourier-transform-fft

Gong, Y., Chung, Y.-A., & Glass, J. (2021). AST: Audio Spectrogram Transformer (arXiv:2104.01778). arXiv. https://doi.org/10.48550/arXiv.2104.01778

Hartmann, W. M. (2013). Principles of Musical Acoustics. Springer Science & Business Media.

Lostanlen, V., Andén, J., & Lagrange, M. (2019). Fourier at the heart of computer music: From harmonic sounds to texture. Comptes Rendus. Physique, 20(5), 461–473. https://doi.org/10.1016/j.crhy.2019.07.005

Lu, W.-T., Wang, J.-C., Won, M., Choi, K., & Song, X. (2021). SpecTNT: a Time-Frequency Transformer for Music Audio.

Medley-solos-DB: A cross-collection dataset for musical instrument recognition. (n.d.). https://doi.org/10.5281/zenodo.1344103

Pellman, S. (1994). An Introduction to the Creation of Electroacoustic Music (1st ed.). Wadsworth Publishing Company.

Phillips, H. (2023, May 14). Positional Encoding. Medium. https://medium.com/@hunter-j-phillips/positional-encoding-7a93db4109e6

Phuong, M., & Hutter, M. (2022). Formal Algorithms for Transformers.

Pipes, M. (2018). A Comparison of Saxophone Mouthpieces Using Fourier Analysis to Quantify Perceived Timbre. Dissertations. https://digscholarship.unco.edu/dissertations/545

Roberts, L. (2022, August 17). Understanding the Mel Spectrogram. Analytics Vidhya. https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

Sperber, M., Niehues, J., Neubig, G., Stüker, S., & Waibel, A. (2018). Self-Attentional Acoustic Models (arXiv:1803.09519). arXiv. https://doi.org/10.48550/arXiv.1803.09519

Sundberg, J., & Nordström, P.-E. (1976). Raised and lowered larynx—The effect on vowel formant frequencies. STL-QPSR, 6, 35–39.

Thilakarathne, H. (2021, July 31). Handling Imbalanced Classes with Weighted Loss in PyTorch. NaadiSpeaks. https://naadispeaks.blog/2021/07/31/handling-imbalanced-classes-with-weighted-loss-in-pytorch/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention Is All You Need (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Velardo, Valerio (Director). (2021, July 1). How to Implement a CNN for Sound Classification. https://www.youtube.com/watch?v=SQ1iIKs190Q

What are Convolutional Neural Networks? | IBM. (n.d.). Retrieved November 24, 2023, from https://www.ibm.com/topics/convolutional-neural-networks

What is acoustic impedance? (n.d.). Retrieved November 22, 2023, from https://www.phys.unsw.edu.au/jw/z.html

What is Timbre in Music? Description and Examples - Hoffman Academy Blog. (n.d.). Hoffman Academy. Retrieved November 21, 2023, from https://www.hoffmanacademy.com/blog/what-is-timbre-in-music-description-and-examples/

Won, M., Choi, K., & Serra, X. (2021). Semi-Supervised Music Tagging Transformer (arXiv:2111.13457). arXiv. http://arxiv.org/abs/2111.13457

Yu, G., & Slotine, J.-J. (2008). Audio Classification from Time-Frequency Texture (arXiv:0809.4501). arXiv. http://arxiv.org/abs/0809.4501

Zhang, J. (2021). Music Feature Extraction and Classification Algorithm Based on Deep Learning. Scientific Programming, 2021, 1651560. https://doi.org/10.1155/2021/1651560