Automatic Instrument Recognition and Music Scripting

Amina Yasser (*Media Engineering Technology*) and **Prof. Mohammed Salem** (*Digital Media Engineering*)

The German university in Cairo

amina.sakr@student.guc.edu.eg

Automatic music transcription (AMT) is the process of converting music signals into some form of music notation. This is an extremely challenging task in the signal processing and artificial intelligence field particularly for music containing multiple simultaneous notes with multiple instruments. Thus, identifying the musical instruments utilized in a polyphonic music recording could enable the extraction of valuable data, which could be crucial in the development of music information retrieval (MIR). This could help with a variety of applications, including music genre classification, recommender systems, and more accurate music transcription.

Literature Review

Lara Haidar (2019) [1] proposes the use of a multi-class classifier in audio to detect the presence of pre-selected instruments. In this paper, the instruments were divided into four categories (Piano, Drums, Flute, or Other). A convolutional neural network is used to classify the data (CNN) . AudioSet is a dataset by Google [2], it includes labeled data in the form of 10-second YouTube clips. Because the dataset may contain background noise and involve several instruments, the processing stage might well be challenging. There are 9600 samples in total, with 2400 examples in each class. There are 8000 samples for training, 800 samples for validation, and 800 samples for evaluation.

The data is pre-processed at 16000 Hz, then downmixed to only use one channel before being normalized. The input for the model is then represented in mel-spectrograms. Before extracting the mel-spectograms, data augmentation techniques are applied to integrate random noise to the samples to minimize overfitting. There are 8000 samples in the training set, 128 batches, 15 epochs, and a learning rate of 10e-3. These values were chosen after several trials. Cross-entropy with Adam optimizer is the loss function used. An early stop was made as soon as the model began to overfit the training data. This was verified by computing the evaluation metrics in the training set and comparing them to the precision and recall metrics in the validation set.

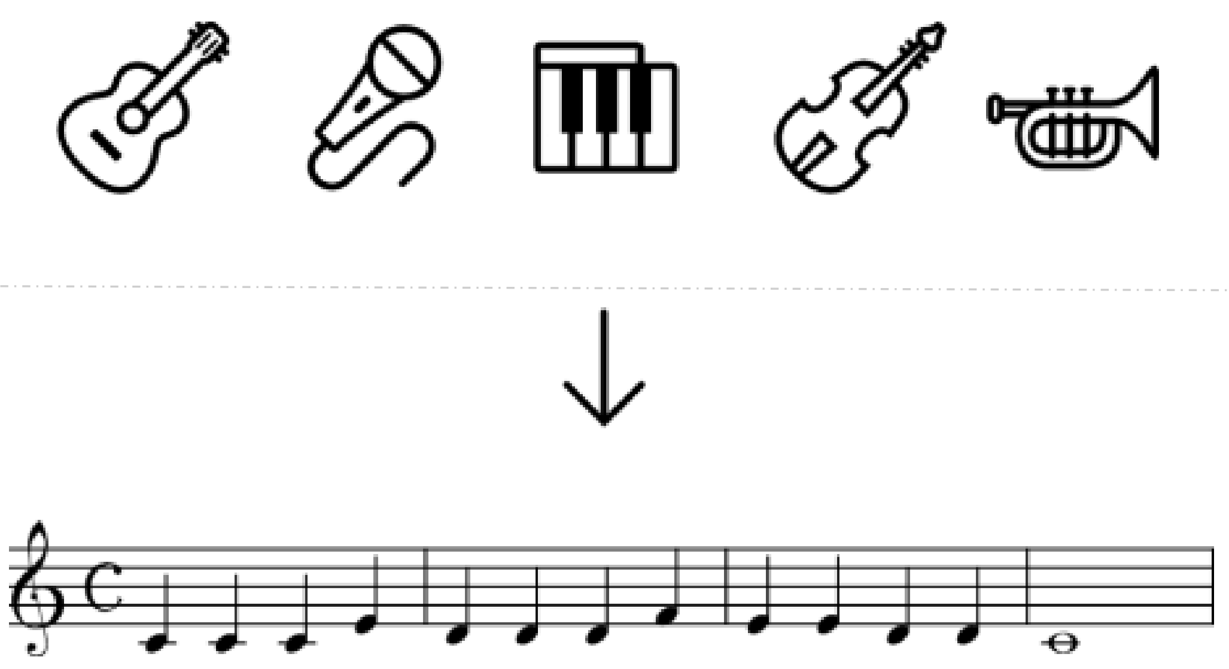
As for the results, when computing the confusion matrix, the largest errors were observed between the classes Drums and Others. Precision, recall, and F1-score were the test set measures. With an average recall of 65 percent and an average F1-score of 64 percent, the model was 70 percent precise. More architectures could be investigated to improve performance, such as RNN [3], RCNN [4], or CRNN [5], which have been shown to work better with audio. More data for the Drums class might also be provided, allowing the model to learn from the characteristics of the drums' sound. Furthermore, classes could be separated into sub-classes; for example, instead of piano, keyboard and classical piano could be introduced, and other instruments could be added.

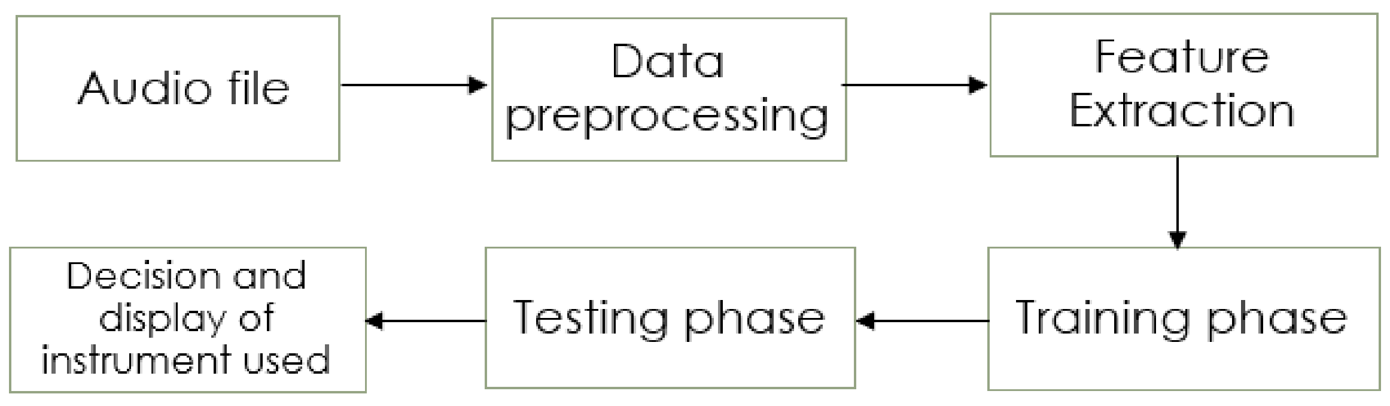
Methodology

For the instruments recognition, the dataset was created manually reaching a total of 37 instruments. The dataset is split into train and test sets, with no overlap between the two sets. Also data augmentation techniques are used for the files in the train set to prevent overfitting. Audio is taken as an input to the preprocessing stage, where it is down-sampled to 22.05KHz and normalized. The necessary features are extracted by extracting the MFCCs for every file in the dataset. This gives distinctive features for each audio file which could then be converted into a data frame and fed to the deep learning model.

For the music transcription, the wav audio file is taken as an input, then the corresponding mid file is generated from the model as well as the piano-roll representation. Then the musical notes are produced from the mid file.

Also, a web application is created to visualize the inputs and outputs.

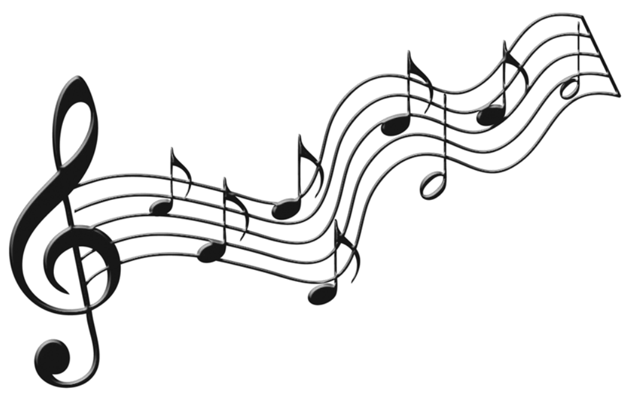




Conclusion

Using deep learning techniques and particularly Convolutional Neural Networks (CNNs), the instruments used in a song or any musical audio whether monophonic or polyphonic are identified and displayed. Also, the audio is transcribed to produce the corresponding musical notes. This could be extremely beneficial in the MIR field.

The system was robust for the instruments recognition, as the train and test accuracies were very close. This means that the system is not overfitting. However, more instruments could be added and more polyphonic music could be added with better quality.



References

1. L. Haidar-Ahmad, "Music and instrument classification using deep learning tech-nics," Recall, vol. 67, no. 37.00, pp. 80–00, 2019.
2. Google, "Audioset, https://research.google.com/audioset," 2019
3. E. C. Kevin Vu, "Recurrent neural networks: Deep learning for sequential data,"2020.Carlsen, L. & Zhou, K. (2012).
4. R. Gandhi, "R-cnn, fast r-cnn, faster r-cnn, yolo—object detection algorithms," towards data science, vol. 9, 2018.
5. C. C. Chatterjee, "An approach towards convolutional recurrent neural networks,"Medium. url: https://towardsdatascience. com/an-approach-towards-convolutional-recurrent-neural-networks-a2e6ce722b19 (visited on 12/28/2019), 2019.
6. Y. Han, J. Kim, and K. Lee, "Deep convolutional neural networks for predominant instrument recognition in polyphonic music," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 1, pp. 208–221, 2016.

E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," IEEE Signal Processing Magazine, vol. 36, no. 1, pp. 20–30, 2018.

A. Lucena, C. Moraes, K. Nose-Filho, D. Fantinato, A. Neves, and R. Suyama, "Musical instruments recognition using machine learning techniques: Mlp and svm," 10 2020Fossen, F. M., & Sorgner, A. (2019).

R. SUNIL, "Understanding support vector machine (svm) algorithm from examples (along with code)," 2017

). C. Bento, "Multilayer perceptron explained with a real-life example and python code: Sentiment analysis," 2021.

Results

After training the dataset, it is shown that given a wav audio file, the instruments used are automatically recognized with the accuracies shown in table 1. Also given an audio file, the corresponding piano roll representation and musical notes are generated as shown in figure 1 and 2. Also, a web app is created to visualize both.

	Monophonic	Polyphonic(2 insts)	Polyphonic(5 insts)	Polyphonic(10 insts)
Training	94%	76%	67%	64%
Testing	90%	75%	65%	62%

Table 1 results for instruments recognition

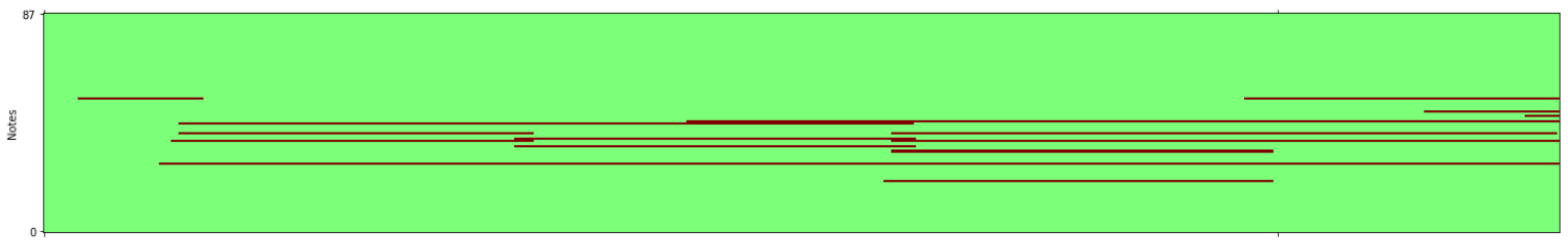


Figure 1 resulting piano roll representation for a given wav file

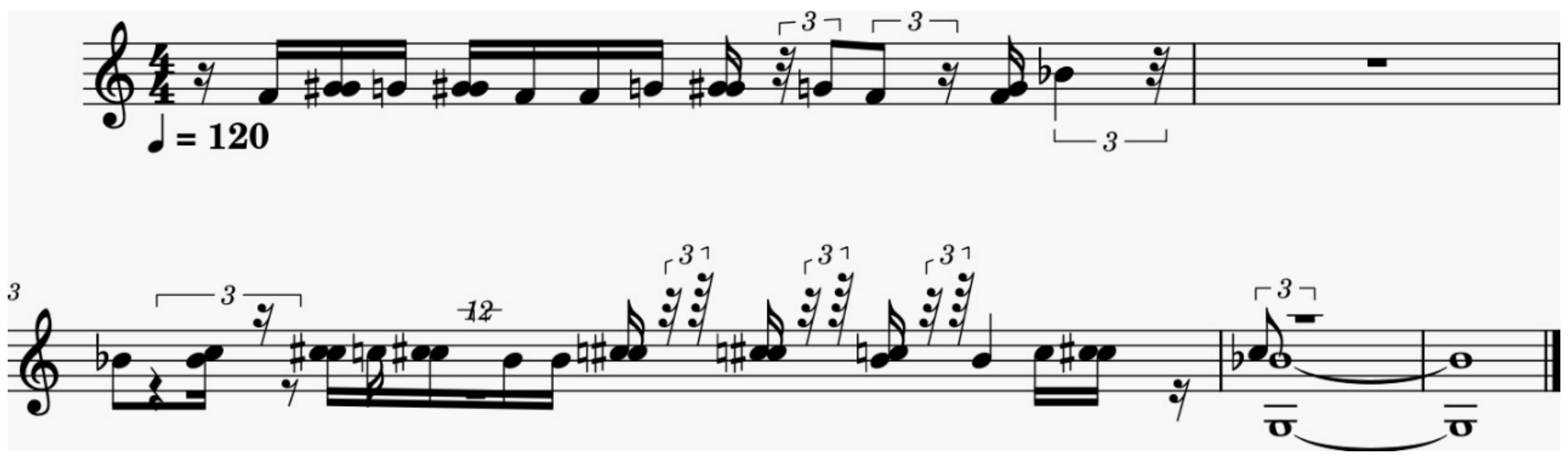


Figure 2 resulting musical notes for a given wav file

Prepared for Thesis Poster Display Conference

11th -12th June 2022