2021 9th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC) | 978-1-6654-8292-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/JAC-ECC54461.2021.9691427

# Audio Signal Processing and Musical Instrument Detection using Deep Learning Techniques

Sally M. Elghamrawy
*Computer Engineering Department*
*Misr Higher Institute for Engineering and technology*
Mansoura, Egypt – Senior IEEE member
*Sally_elghamrawy@ieee.org, ORCID: 0000-0002-5430-390X*

Shehab Edin Ibrahim
*Computer Engineering Department*
*British University in Egypt*
Cairo, Egypt

**Abstract- The advance of deep learning and audio signal processing techniques has led to serious development on Musical Information retrieval (MIR). Effective audio processing can improve speed, reduce errors, and sometimes increase the accuracy of detecting musical instrument. Spectrographic data is also necessary for many mathematical tools common across Musical Information retrieval. A major aspect of MIR is the categorization of pieces of music. One of the main tools used for categorization tasks in recent years is deep learning, which has led to many advancements in MIR. One such categorization task that deep learning is useful for is the recognition of instruments in a piece of music. In this paper, a new architecture is proposed for audio processing and musical instrument detection using Multilayer Perceptron (MLPs), Convolution Neural Networks (CNN), and Recurrent Neural Networks - Long Short Term Memory (RNN-LSTM). In addition, a number of experiments are implemented using real dataset that contains 20,000 recording. The three deep learning techniques are implemented and compared to present potential new solutions. The usage of processing techniques unique to the field of deep learning is also discussed.**

***Keywords— Deep Learning, Convolution Neural Networks, audio signal processing, Musical Instrument Detection.***

## I. INTRODUCTION

Musical Information Retrieval is a new field of research, its purpose is extracting data and metadata from musical recordings. It brings together the fields of machine learning, digital signal processing(DSP), and musicology [1]. MIR includes some relatively simple tasks such as extracting the BPM (Beats Per Minute) of a musical track. Other properties of music are more difficult to determine, such as recognizing human singing voices. Machine learning is an effective tool when it comes to complex categorization tasks. Previous research has shown its usefulness when it comes to MIR [2]. One such complex categorization task is identifying instruments that are present in a recording, especially when more than one instrument is present. Deep learning presents an interesting solution to the problem of instrument recognition [3]. Neural networks (NNs) are a type of deep learning technique that emulates the way biological brains learn. Large quantities of data are used to train neural networks [4]. This paper will discuss three different kinds of neural networks. These are Multilayer Perceptron (MLPs) [5], Convolution Neural Networks (CNNs) [6], and recurrent neural networks (RNNs). Previous research in instrument recognition has shown that the type of network used has a large impact on outcome. One of the goals of this paper will be to produce a comparison of these three different neural network architectures.

### *A. The Challenges in instrument detection*

The first challenge of using a neural network for instrument detection is using a diverse dataset that represents a large set of instruments. The size and diversity of the dataset is particularly important when it comes to detecting multiple instruments. The second challenge in instrument detection is the difficulty of detecting multiple instruments that are present in the same track. most music uses multiple instrument, especially since the human voice is considered and instrument for our purposes. Being able to detect multiple instrument is a desirable feature for any Instrument recognition program. Therefore, datasets specifically suited for multiple instrument tasks need to be developed. Datasets suitable for multiple instrument tasks would have more complexity than single instrument datasets. [7]. Multiple instrument detection would also require more mature neural network architectures and audio processing techniques.

### *B. Audio processing*

The pre-processing of audio data is an important first step in most MIR tasks. Effective audio processing can improve speed, reduce errors, and sometimes increase accuracy. Spectrographic data is also necessary for many mathematical tools common across MIR. This means that effective audio content processing is an important topic of questioning in this paper. Deep learning techniques also present new methods of audio content processing. For example, auto encoding neural networks that are used for dimensionality reduction by ignoring insignificant data.

### *C. The Paper's main contribution*

Explore the usefulness of RNN-LSTM networks for instrument recognition. This is to contrast with the large body of research that uses CNNs. Investigate different multi-instrument recognition techniques. Previous research has largely dealt with single instrument detection. Most datasets are also single instrument which may have guided most research in the past. Using the openMic-2018 dataset [7], as it is one of the largest multiple instrument dataset that has not seen much representation in research. Using the Voguish deep learning auto-encoder. The VGGish [8] auto encoding process was developed based on auto encoding methods used for computer vision. It is a relatively new method for audio processing, and is especially suited for deep learning tasks. The rest of the paper is organized as follows: section II summarizes the recent work in MID. The main proposal is presented in section III. The experimental results are shown in section IV. Finally, the paper is concluded in section V.

## II. RELATED WORK

The first step in the literature review was to determine which types of neural network or deep learning technique the researcher employed and how it influenced the study's outcomes. While isolating the precise influence of the neural network architecture or technique utilized can be difficult (differences in outcome can be attributed to a variety of factors, including dataset used and audio processing methodology.)If a highly effective type of neural network can be established, it might be very valuable and save a lot of time. Convolutional neural networks, or CNNs, are probably the most frequent type of neural network studied in the past [9-11,3]. It's not as straightforward as just employing a standard CNN; several different types of convolutional neural networks

have been used by different academics in their studies. Solanki and Pandey employed AlexNet, a CNN with expanded capabilities that allows it to better work with spectrograms and music, in their 2019 publication. This is especially relevant because CNNs are commonly used to classify or recognize objects in photos. AlexNet's particular advantages make it a possible CNN choice [9]. Convolutional recurrent Neural Networks, or CRNNs, are another type of CNN that has been used in musical instrument recognition research. Gururani et al. made excellent use of a CRNN. A CRNN is a type of CNN that incorporates recurrent neural network features. According to the paper, this allowed the neural network to learn more effectively from data evolution over time. It aided learning from the evolution of spectral data in particular [12]. In prior studies on the area of instrument recognition, more standard CNN architectures are also present. Mimilakis et al. utilized a CNN-based technique for instrument recognition for western classical music. Their CNN-based technique produced an impressively accurate system, albeit the authors emphasize the need of effective audio processing in reaching this accuracy [11]. MLPs have a place in the literature as well. Toghiani et al. use an MLP to develop an instrument recognition system in their study. Their research is intriguing not just because of the novel use of an MLP, but also because they set out to achieve recognition using limited time or frequency frames, such as only the first several seconds of a sound or the first 100Hz. This is owing to their attempt to better model human recognition abilities, as earlier study has shown that for a human listener, simply the beginning of an instrument's sound, referred to as the attack, is enough for correct instrument identification. This is a fascinating area for investigation. Their work is especially noteworthy for its emphasis on back propagation as a means of the final models accuracy [13].

It's worth mentioning, primarily as a historical footnote, that early research into applying machine intelligence for instrument recognition predates the widespread use of deep learning techniques in research. A KNN (K closest neighbor) technique was employed in two fundamental works in instrument recognition. While the overall precision and computational efficiency of the models were limited in this era, these publications were significant in identifying contemporary research avenues such as the potential of detecting instruments using the attack [10]. Few prior studies have used RNNs; aside from one implementation of a CRNN by one team of researchers, this form of neural network has seen very little use in the field of instrument recognition. This is noteworthy because RNNs are one of the most commonly used types of neural networks for retrieving musical information. Or, in general, any audio processing task[5].This apparent hole in methodology is intriguing, and it's a possible area where this paper might contribute significant value to instrument recognition research. The use of LSTM (Long Short Term Memory) elements is another feature that is common in musical information retrieval tasks. The applicability of RNN and LSTM usage in Instrument Recognition will be further investigated in this article.

## III. The Proposed Architecture

The proposed architecture contains three main layers as in figure 1. This architecture is very general to accommodate the different types of neural networks that will be tested. Each network has a more specific design. The models first layer is the pre-processing layer. The first purpose of this layer is to optimize the data, firstly by ensuring that there are no redundancies and also splitting audio files into smaller files that are a more appropriate as input into the next layer. This optimization stage is important to reduce training time down the line.

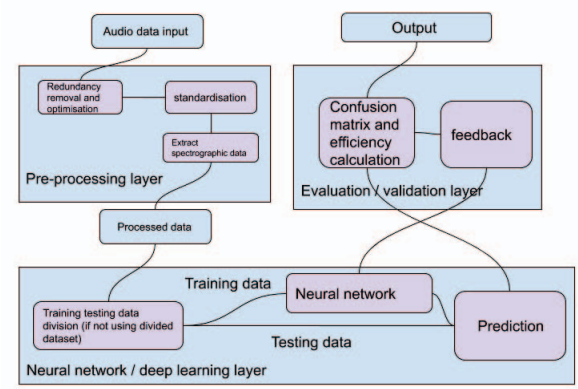


Fig. 1. The Proposed architecture

The following step is to generate spectrographic and frequency domain data from the audio. Spectrographic representations of audio data are necessary for feature extraction and auto encoder. Spectrographic data is therefore necessary for many parts of the model. The literature often stresses that effective pre-processing is an integral part of improving model precision. Spectrographic processing can be further separated into a Short Time Fourier Transform(STFT) [14], The output of which is then process into Mel Frequency Cepstrum (MFC)[15]. The output of the MFC is passed onto the VGGish auto-encoder which serves to reduce the dimensionality of data and extract features. The output of the preprocessing layer is used the second layer's input. The second layer of the architecture is the neural network layer. The first step of which is separating the input data into a training dataset and a testing dataset. The OpenMic-2018 dataset project team provide a training test split with deep learning in mind. This split is prepared to ensure a diversity in artist, genres and instruments across the training and testing sets. Other datasets also have similar pre generated splits, for example the IRMAS dataset for single instruments. The training set output of the split is passed onto the training stage.

This is the step where the neural network develops, when training is complete the neural network can begin generating labels based on the test dataset. This represents the output of the neural network and these predictions are passed on as the input to the final layer. The final layer's task is evaluation and validation. accuracy metrics are used to compare the prediction of the neural network with the testing data. accuracy and error values are calculated at this stage. Feedback is then back propagated to the neural network. This stage is crucial to neural network's ability to learn. Once the Network is trained and tested its output is passed on to the validation layer.

### A. *Technical approaches used*

Three neural network models are tested in this paper. The first one is the MLP neural network. Its primary purpose is producing baseline results that can be used to assess the other networks. The second is a CNN. The design of which similar to the AlexNet architecture which has been used previous research. The CNN will serve as a real test for the unproven RNN-LSTM design. This will also serve to inform us on how existing methods act when dealing with multiple instrument detection tasks. The last network is a RNN-LSTM (Recurrent Neural Network, Long Short Term Memory) RNN-LSTM neural network architecture is a new approach when it comes to instrument recognition, but has been shown effectiveness in other audio classification tasks. Python is used as ssignal processing is straight forward in python with many well developed libraries. The VGGish auto encoder was also developed for use in python. Some of the libraries used are the

tensor flow libraries, Keras, and Sci-kit learn. The open-mic2018 dataset has built in example scripts in python and a python library.

### B. *Pre-processing*

The pre-processing step is essential for the neural network's proper performance. The major goals of the pre-processing stage, as previously stated, are to present inputs to the network that are more reliable and result in improved accuracy. These tasks will be carried out with the help of high-level Python libraries.

### C. *Data Normalization*

Normalization is a crucial stage in the audio data processing process. Specifically, all recordings should be within the same amplitude range, sampling rate, and length. Because the OpenMic-2018 has already been normalized, this step can be mostly ignored [7].

### D. *Generation of STFT*

The STFT is the initial stage in creating the model's robust spectrographic data. The STFT is a set of Fast Fourier transforms performed at a predetermined sample interval.

A relevant representation of how the spectrographic data evolves over time is obtained from the STFT stage [16].

Librosa is used to generate the STFT for each signal. Librosa is a python package for high-level audio processing.

From an audio signal, the following function will generate an STFT: `librosa.core.stft(signal, hop_length, n_fft)`

The hop length parameter specifies the number of samples that each FFT will examine.

The FFT window size is described by n fft.

### E. *Generation of MFC*

The second step in the process is to create a Mel Frequency cepstrum (MFC).The MFC is also a depiction of how an audio signal's spectral composition evolves over time.

Beyond that, it's a good representation of power spectral density that's meant to mimic human hearing more precisely. Because of its power, the MFC is commonly used in MIR.

The MFCC can be constructed using Librosa in the same way as the STFT:

```
librosa.feature.mfcc(signal, n_fft, hop_length, n_mfcc)
```

Except for the addition of n mfcc, which corresponds to the number of MFCC components generated, which specify the resolution of the final MFC, the syntax is quite similar to that of STFT creation.

### F. *VGGish auto encoding*

Dimensionality reduction and feature extraction using the VGGish auto encoder are the final key steps in the processing layer. VGGish is a Google-developed auto encoder. VGGish provides spectrographic data that is very compressible when compared to an MFC or other standard type of spectrographic data. It was developed utilizing a deep learning approach, based on the VGG architecture for auto encoding of image data. It is worth noting that VGGish is itself a neural network that has been pre-trained by its developers on audio data.

## IV. The Experimental Results

The performance of each deep learning technique used in the model is displayed in the following graphs. The first metric is accuracy, it tests the model's accuracy in single instrument detection. The second metric used is Top K categorical accuracy, it shows us the accuracy of the model when it comes to detecting multiple instruments, the value of k used is five. Meaning that it calculates the accuracy of the models top five predictions for each recording. Loss is the third metric used it helps in understanding the overall effectiveness of the model used. To calculate the models overall ability to distinguish between weak and strong negative labels we use categorical accuracy. Categorical accuracy measures the accuracy of the model when it comes to all 20 instrument labels. It is consistently low when compared to single instrument accuracy and top k categorical accuracy. The following table Shows the properties of the RNN-LSTM, CNN and MLP networks used.

TABLE 1: THE PROPERTIES OF THE RNN, CNN AND MLP NETWORKS USED

| | RNN-LSTM | CNN | MLP |
|---|---|---|---|
| No. of hidden Layers | 4 | 5 | 3 |
| Dimensions | (128*128)*128*64 | (34*64*128*256)*256 | 512/256/64 |
| Max Pooling | - | 3*3 | - |
| Loss function | Binary Cross Entropy | | Categorical cross entropy |
| Activation function of output layer | Sigmoid | | |
| Neuron Drop out | 0.25 | | |
| Learning rate | 0.0001 | | |
| Optimizer | Adam [17] | | |

### A. *MLP*

The highest single instrument accuracy reached by the MLP model is 35%. the loss is 29.95 which is an incredibly high loss.

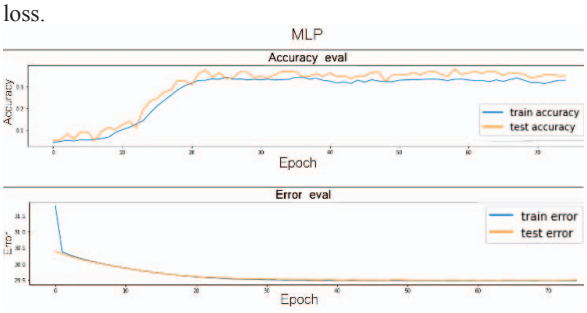


Fig. 2. MLP Results

This is the expected out of the MLP model as it is a naive approach. The MLP model's purpose is to act as a sanity test for the architecture and ensure its functionality, and to act as a simple baseline to compare the effectiveness of the other models. We can see that test and Training results stay close to each other throughout the training period. Both accuracy and categorical accuracy test results tend to vary slightly across epochs but training accuracy reaches a platau and stays at it. The single advantage of the MLP model when compared to the other two models is the training time. Each generation of the MLP model took around 3 seconds to train which is less than 10 percent of the time taken for the other models to train.

### B. *CNN*

The CNN architecture used here is based on the AlexNet model. Immediately it can be seen that the accuracy achieved is significantly higher than the one used by the MLP model. The Single instrument accuracy achieved is 91.5%. The state of the art at the time of writing is 92% for single instrument detection using CNNS. Showing that the approach proposed by this paper achieves results comparable to other research in the field. Multiple instrument recognition as expected has a lower accuracy. Achieving a maximum test accuracy of 84 percent. For multiple instrument detection the highest accuracy reached is 84% with a loss of 0.204. We can already see that the CNN produces some pretty impressive results. With a higher multiple instrument accuracy than initially expected. The CNN model used also consistently showed significant divergence between test and training accuracy as the epochs progressed. This is a clear sign of overtraining that was persistent across changes to the test/train ratio used, and layer sizes/ architecture. The issue of overtraining was harder overcome in the CNN network and is clearly the main issue standing in the way of improving the model. The CNN model also had the highest training time per epoch. With the average epoch time being 37 seconds.

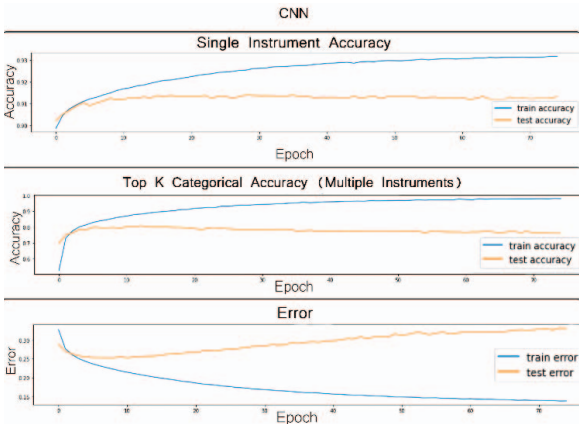


Fig. 3. CNN Results

### C. *RNN-LSTM*

As for the RNN-LSTM network we see an overall improvement in results across the board. The single instrument accuracy is the highest of all three models with the maximum single instrument reaching 94%.This is comparable to or higher than the state of the art when it comes to single instrument detection. The most relevant performance increase when it comes to the goals of this paper is the higher multi-instrument accuracy. The highest reached here is 86% which is two percent higher the highest accuracy reached by the CNN model. It is possible that this number will be improved in future research. The loss is also the lowest, with a loss value of around 0.175. The highest categorical accuracy reached is 52%. Categorical accuracy is used to assess the RNNs ability to differentiate between strong and weak negative labels. Another advantage that the RNN-LSTM architecture provides is that it does not reach a state of over-fitting as quickly as the CNN model. The training and testing accuracies maintain a steady pace with each other for most of the training period and we begin to see overtraining only in the last few epochs. This network also has lower training times than the CNN, with an average time of 29 seconds per epoch.

TABLE 2: THE NETWORK TYPE ACCURACY OF THE RNN, CNN AND MLP

| Network Type | MLP | CNN | RNN-LSTM |
|---|---|---|---|
| Single Instrument Accuracy | 35% | 91.5% | 94% |
| Top k Categorical accuracy | - | 84% | 86% |
| Categorical Accuracy | - | - | 52% |
| Loss | 29.95 | 0.204 | 0.175 |

## V. CONCLUSION

Instrument detection and audio content processing are interesting problems which have compounding effects on the rest of MIR. They are also problems which present many interesting challenges, which are uniquely suited for novel deep learning techniques. This paper has achieved measurable success in its goal of evaluation the usefulness of RNN-LSTM neural networks and Deep learning auto encoders when it comes to multi instrument detection. With both of these new approaches producing promising results. There are many potential avenues for new research based on these results. Future research could experiment with new network architectures, or variations of the RNN architecture presented here, as well as further experimentation with deep learning tools for signal processing before the deep learning stage. There is great potential for research achieving multi instrument detection accuracy of above 90% in the near future.

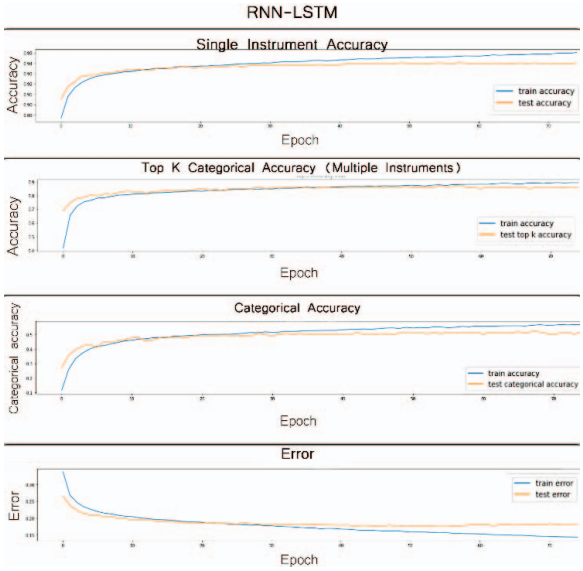


Fig. 4. RNN-LSTM

## REFERENCES

[1] Downie, J. S. (2003). Music information retrieval. Annual review of information science and technology, 37(1), 295-340.

[2] Patterson, G., Pfalz, A., & Allison, J. (2017). Neural Audio: Music Information Retrieval Using Deep Neural Networks.

[3] Gómez, J. S., Abeßer, J., & Cano, E. (2018). Jazz Solo Instrument Classification with Convolutional Neural Networks, Source Separation, and Transfer Learning. In ISMIR (pp. 577-584).

[4] Choi, K., Fazekas, G., Cho, K., & Sandler, M. (2017). A tutorial on deep learning for music information retrieval. arXiv preprint arXiv:1709.04396.

[5] Lostanlen, V., Andén, J., & Lagrange, M. (2018, September). Extended playing techniques: the next milestone in musical instrument recognition. In Proceedings of the 5th International Conference on Digital Libraries for Musicology (pp. 1-10).

[6] Elghamrawy, S.M., Hassnien, A.E. and Snasel, V., 2021. Optimized deep learning-inspired model for the diagnosis and prediction of COVID-19. Cmc-Computers Materials & Continua, pp.2353-2371.

[7] Humphrey, E., Durand, S., & McFee, B. (2018, September). OpenMIC-2018: An Open Data-set for Multiple Instrument Recognition. In ISMIR (pp. 438-444).

[8] Shi, L., Du, K., Zhang, C., Ma, H. and Yan, W., 2019. Lung sound recognition algorithm based on VGGISH-bigru. IEEE Access, 7, pp.139438-139449.

[9] Solanki, A., & Pandey, S. (2019). Music instrument recognition using deep convolutional neural networks. International Journal of Information Technology, 1-10.

[10] 2Siedenburg, K., Schädler, M. R., & Hülsmeier, D. (2019). Modeling the onset advantage in musical instrument recognition. The Journal of the Acoustical Society of America, 146(6), EL523-EL529.

[11] Taenzer, M., Abeßer, J., Mimilakis, S. I., Weiß, C., Müller, M., Lukashevich, H., & Fraunhofer, I. D. M. T. (2019). Investigating cnn-based instrument family recognition for western classical music recordings. ISMIR.

[12] Gururani, S., Summers, C., & Lerch, A. (2018, September). Instrument Activity Detection in Polyphonic Music using Deep Neural Networks. In ISMIR (pp. 569-576).

[13] Toghiani-Rizi, B., & Windmark, M. (2017). Musical instrument recognition using their distinctive characteristics in artificial neural networks. arXiv preprint arXiv:1705.04971.

[14] Wang, L.H., Zhao, X.P., Wu, J.X., Xie, Y.Y. and Zhang, Y.H., 2017. Motor fault diagnosis based on short-time Fourier transform and convolutional neural network. Chinese Journal of Mechanical Engineering, 30(6), pp.1357-1368.

[15] Logan, B., 2000. Mel frequency cepstral coefficients for music modeling. In International Symposium on Music Information Retrieval.

[16] Yan, P.Z., Wang, F., Kwok, N., Allen, B.B., Keros, S. and Grinspan, Z., 2019. Automated spectrographic seizure detection using convolutional neural networks. Seizure, 71, pp.124-131.

[17] Elghamrawy, S.M., Hassanien, A.E. and Vasilakos, A.V., 2021. Genetic-based adaptive momentum estimation for predicting mortality risk factors for COVID-19 patients using deep learning. International Journal of Imaging Systems and Technology.