2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM)
Amity University

# Music Instrument Recognition using Machine Learning Algorithms

Shreevathsa P K
Department of E. C. E
PES University
Bengaluru, India
vathshree@gmail.com

Harshith M
Department of E. C. E
PES University
Bengaluru, India
mharshith@gmail.com

Abhishek Rao MDepartment of E. C. E
PES University
Bengaluru, India
abhishekraom9@gmail.com

Ashwini
Department of E. C. E
PES University
Bengaluru, India
ashwinib@pes.edu

***Abstract*—In this is modern era, everyone listens to and plays music. Music is diverse across the globe. It is a language that speaks by itself and is the fulcrum of all the arts. We can say that rich legacy of this flawless art is infinity and beyond. It has the ability to mesmerise one and all. If there was a way in which we could get to know the instruments that are being played in the music, it would be more interesting.So, we can classify the music based on certain instruments.In last two decades, researchers are actively associated with human perception towards the study of Musical Instruments. Our work is based on designing an application that recognizes the instruments that are present in a given Music. It is no hidden fact anymore that neural network in general are dominating every single aspect and application in this world. The fact that a neural network can do a task that can be done by a human brain efficiently if we train them properly tells us a lot about their advancement in the day to day world. Thus, we make use of neural networks for the recognition of an instrument in a piece of music. Hence, we make use of ANN (Artificial Neural Network) and CNN(Convolutional Neural Network). We've considered eight different music instruments. The models thus built play a major role in finding the music instruments being played. Various features are extracted in the process and the result is found.**

***Keywords—Instrument Recognition, MFCC, CNN, ANN***

## I. INTRODUCTION

Music instrument recognition is a fascinating and essential thing in music indexing, retrieval and automatic transcription.Machine learning and Deep learning are great tools for music application and analysis. ML and DL are certainly helping the new music listeners to improve their knowledge of music and also aid them in recognizing the instruments properly. We think about the marvels of ML and DL and the disruptions it has brought into the music industry. It is also a good tool for polyphonic pitch tracking and is used in real time music performance. The task in ML is essentially to extract features from the audio. Extraction of features helps in determining which instrument is being played or which note is being played or even further. But, the drawback of ML is that we need to know what the best features are. This is where DL plays a significant role without extracting the features. DL lets the algorithm to figure it out. However, despite of all the features and methods of classification, there is still a lot of subjectivity based on which method can be used to classify a particular piece of music. Thus, in this paper, we implement models using both ML (ANN) and DL (CNN). In ML, certain features like cepstral coefficients and zero crossing rate are used. In DL, the model is directly designed without having to extract any features and thus directly meet the requirements.

## II ARTIFICIAL NEURAL NETWORK (ANN)

ANN is basically machine learning. ANN is based on a collection of connected units or nodes called artificial neurons. They're inspired by, but not similar to biological neural network. There are different layers i.e., Input layer, Hidden layers and Output layers. Each network is assigned a weight, which has relative importance. ANN has gained a lot of importance in the recent past. Most of the modern-day applications are being done using ANN models. A propagation function or an activation function is used in determining the output of the network. They deliver outputs



978-1-7281-0666-3/20/$31.00©2020IEEE

based on predefined activation function. It provides a flexible way to handle regressions and classification problems without the need to explicitly specify any relationship between input and the output. In the neural network we have used, we extract certain features required to recognize the instruments in the music. We use ZCR and MFCC of different instruments to determine the instruments being played.

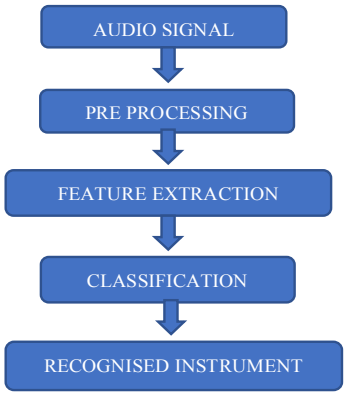



Fig. 1. Block Diagram of Artificial Neural Network

## III FEATURE EXTRACTION

In the ANN model, features must be extracted to determine the instrument used. Here, we've used Mel-Frequency Cepstral Coefficients (MFCC) and Zero Crossing Rate (ZCR). ZCR is pretty commonly used in the most of the audio processing techniques. But MFCC especially separates different classes of music instruments. These two features combined gives various characteristics of the music piece being played and helps to recognize the instruments present in it.

### A) MEL-FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

MFCC is a set of features of a music instrument that is uniquely used to characterize the instruments. This precisely describe the overall shape of the spectral envelope. It models the characteristics of a human voice. Generally, humans are much better in recognizing the variation in low pitch compared to the tone with high pitch. Mel tone relates a pitch or a pure tone or to its actual measured frequency.
The formula for converting frequency to Mel scale is given by:

$$M(f) = 1125\ln(1+f/700)$$

Through this formula, the actual frequency is converted to Mel scale which helps in determining its coefficients and hence contributing to the instrument recognition. It is also a power representation, linear cosine transform representation, and of a log power spectrum.

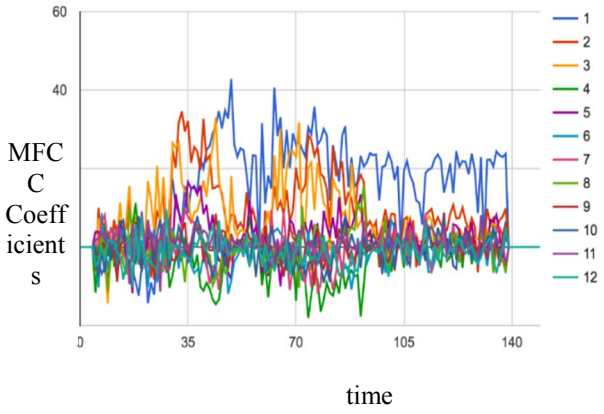



Fig. 2. MFCC graph

The x-axis in the MFCC graph denotes time and the y-axis denotes Mel cepstral coefficients of any given instrument at that particular instant of time.

### B) ZERO CROSSING RATE (ZCR)

ZCR is the rate of sign changes of a signal. If the signal is moving from negative to positive, it is called Positive Zero Crossing and if the signal is moving from positive to negative, the crossing on the axis is called Negative Zero Crossing. This feature can be broadly used in Speech recognition and music instrument retrieval. The following formula is used to determine the Zero Crossing Rate in a typical audio frame.

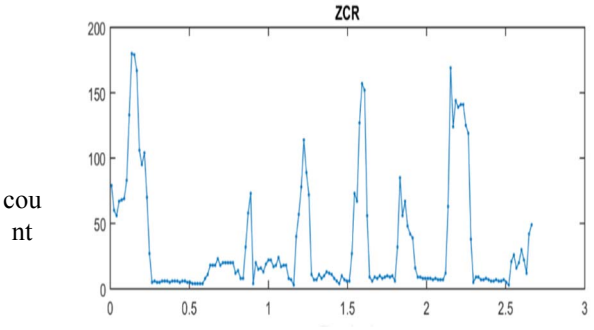



Fig. 3. ZCR graph

The graph above determines the ZCR of any particular instrument. In this graph, x-axis denotes time and y-axis denotes count i.e., zero crossing rate of the instrument at that instance.

## IV.CONVOLUTIONAL NEURAL NETWORK (CNN)

CNN is a type of neural network where all the input layers are not connected to the hidden layers and all the hidden layers are not connected to the output layer. Basically, it is a partially connected layer of neural network. As the name itself indicates, the mathematical operation performed here is convolution. They use convolution instead of matrix multiplication in one of those layers.There are several architectures that are available and have been key in developing algorithms. They are also called Shift Invariant Artificial Neural Network (SIANN). CNNs are the regularized version of a multi-layer perceptron.The first layer which is used to extract features from an image is called Convolution layer. The spectrum obtained from audio signal is given to the convolution layer. Next, there's a flattened layer which provides proper connection strength and takes decision for classification between convolutional layer and fully connected layer. It has one or more convolution layer and also one or more fully connected layer. Pooling layers are also present which are used to decrease the computational power required to process the data. Pooling layers can be max pooling, min pooling, average pooling and other relevantly used pooling layers.This sometimes is also called or used as flattened layer. The matrix from the convolutional layer is formed as vector in flattened layer, as a flattened output and is fed into fully connected layer as a feed forward neural network.The following figure shows the block diagram for a CNN model.

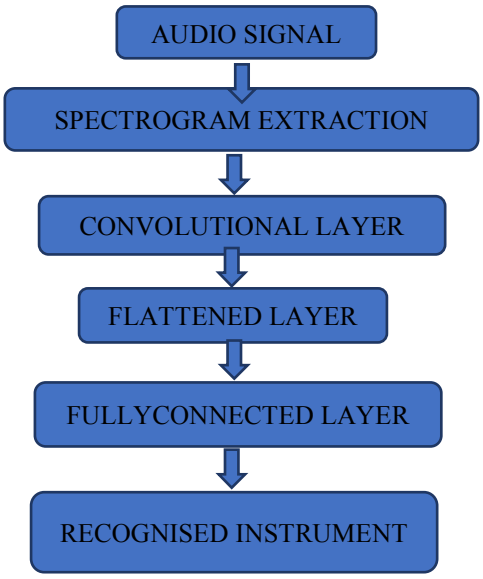



Fig. 4. Block diagram for Convolution Neural Network

## V. RESULTS AND COMPARISONS

The confusion matrices for ANN and CNN are obtained. We've considered eight instruments to determine the accuracy. More the number of samples, more the accuracy. Thus, if the number of samples that are trained are more, more is the possibility of getting the recognition of instrument right. We've used Librosa and Keras library of Python to come up with these two kinds of neural network for the instrument recognition.Here, around 4000 samples are taken for each instrument.

We've obtained the following confusion matrix from these samples. The diagonal values of the confusion matrix give the accuracy of the corresponding instruments. Thus, the accuracy for the instruments are as follows:

| Instrument Name | ANN Accuracy(%) | CNN Accuracy (%) |
|---|---|---|
| Bass Acoustic | 90 | 98.21 |
| Bass Electronic | 73.18 | 94.69 |
| Bass Synthetic | 97.43 | 93.02 |
| Flute | 68.75 | 93.75 |
| Guitar Acoustic | 41.66 | 84.31 |
| Guitar Electronic | 43.18 | 100 |
| Keyboard Acoustic | 23.52 | 72.72 |
| Keyboard Electric | 83.83 | 93.93 |
| Keyboard Synthetic | 90.47 | 100 |
| Mallet | 55.31 | 92.68 |
| Organ | 84.37 | 97.97 |
| Reed | 87.03 | 71.69 |

| String | 81.96 | 98.43 |
|---|---|---|
| Vocal | 88.46 | 100 |

The confusion matrix obtained for ANN

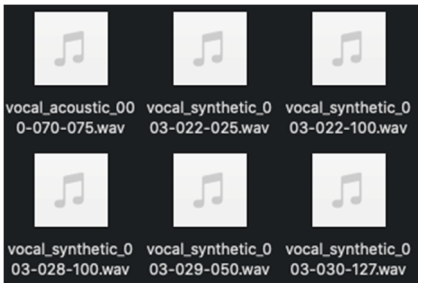
| INSTRU | (I) | (II) | (III) | (IV) | (V) | (VI) | (VII) | (VIII) | (IX) | (X) | (XI) | (XII) | (XIII) | (XIV) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (I) | 90 | 0 | 0 | 3.12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6.55 | 0 |
| (II) | 2.00 | 7318 | 0 | 0 | 1.04 | 2.27 | 0 | 21.21 | 0 | 17.02 | 2.08 | 0 | 0 | 11.53 |
| (III) | 0 | 0.72 | 97.43 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| (IV) | 2.00 | 0.72 | 0 | 68.75 | 5.20 | 0 | 0 | 0 | 0 | 0 | 1.04 | 0 | 1.63 | 3.84 |
| (V) | 2.00 | 7.97 | 0 | 0 | 41.66 | 18.18 | 5.88 | 22.22 | 9.52 | 14.89 | 3.12 | 0 | 0 | 3.84 |
| (VI) | 0 | 0.72 | 0 | 0 | 6.25 | 43.18 | 11.76 | 9.09 | 0 | 6.38 | 4.16 | 0 | 0 | 0 |
| (VII) | 0 | 1.44 | 0 | 0 | 0 | 0 | 23.52 | 10.10 | 0 | 0 | 1.04 | 0 | 0 | 0 |
| (VIII) | 2.00 | 3.62 | 0 | 0 | 4.16 | 0 | 9.09 | 83.83 | 0 | 0 | 2.08 | 0 | 0 | 0 |
| (IX) | 0 | 0.72 | 0 | 0 | 1.04 | 0 | 0 | 0 | 90.47 | 0 | 0 | 0 | 0 | 0 |
| (X) | 2.00 | 4.34 | 2.56 | 0 | 5.20 | 0 | 0 | 7.07 | 0 | 55.31 | 1.04 | 0 | 0 | 0 |
| (XI) | 0 | 0.72 | 0 | 12.5 | 1.04 | 2.27 | 0 | 2.02 | 0 | 0 | 84.37 | 3.70 | 0 | 15.38 |
| (XII) | 6.00 | 0 | 0 | 9.37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 87.03 | 0 | 0 |
| (XIII) | 0 | 2.00 | 2.17 | 0 | 0 | 1.04 | 0 | 0 | 4.04 | 4.76 | 0 | 0 | 81.96 | 3.84 |
| (XIV) | 2.00 | 0.72 | 0 | 0 | 1.04 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 88.46 |

The confusion matrix obtained for CNN

| INSTRU | (I) | (II) | (III) | (IV) | (V) | (VI) | (VII) | (VIII) | (IX) | (X) | (XI) | (XII) | (XIII) | (XIV) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (I) | 98.21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.01 | 0 | 0 | 0 |
| (II) | 0 | 94.69 | 0 | 0 | 1.96 | 0 | 0 | 5.05 | 0 | 0 | 0 | 0 | 0 | 0 |
| (III) | 0 | 0 | 93.02 | 0 | 0 | 2.63 | 0 | 2.02 | 0 | 0 | 0 | 0 | 0 | 0 |
| (IV) | 3.57 | 0 | 0 | 93.75 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| (V) | 0 | 4.54 | 0 | 0 | 84.31 | 0 | 0 | 6.06 | 0 | 2.43 | 2.02 | 0 | 0 | 4.76 |
| (VI) | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| (VII) | 0 | 1.51 | 0 | 0 | 0 | 0 | 72.72 | 2.02 | 0 | 0 | 2.02 | 0 | 0 | 0 |
| (VIII) | 1.78 | 0 | 0 | 0 | 0.98 | 0 | 4.54 | 93.93 | 0 | 0 | 2.02 | 1.56 | 0 | 0 |
| (IX) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 |
| (X) | 0 | 2.27 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 92.68 | 0 | 0 | 0 | 0 |
| (XI) | 0 | 0 | 0 | 3.12 | 0 | 0 | 0 | 0 | 0 | 0 | 97.97 | 0 | 1.56 | 0 |
| (XII) | 19.64 | 0.75 | 0 | 3.12 | 0 | 0 | 0 | 1.01 | 0 | 0 | 0 | 71.69 | 1.56 | 0 |
| (XIII) | 1.78 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 98.43 | 0 |
| (XIV) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |

It is evident from the above confusion matrix that CNN is relatively better compared to ANN since it gives better accuracy. Also, increasing the samples gives even better accuracy when CNN is used.
The following pictures tells us about the tested result of the model where the given audio sample is recognized by the neural network using ANN and CNN.

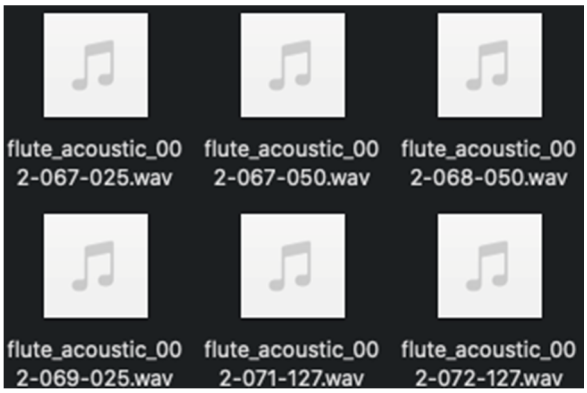
These are few examples of the instruments we've used.




The above image shows few of the samples of vocal acoustic being played.

The following piece of code in the image shows the predictions made for each instrument in general and in this case, prediction made is sarod which is the sample given. The output for the respective instruments is also shown.

The samples given to Flute are:

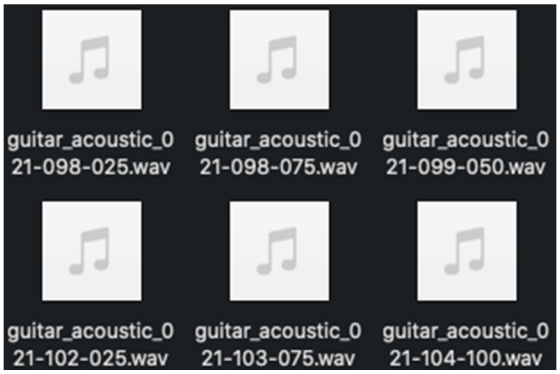



The output is as shown below.

```
print('The instruments given for predicitons are:')
for i in Label_Data_Test[2:3]:#predictions:
    if(i==0):
        print('Bass_Acoustic')
    elif(i==1):
        print('Bass_electronic')
    elif(i==2):
        print('Bass_Synthetic')
    elif(i==3):
        print('Flute')
    elif(i==4):
        print('Guitar_Acoustic')
    elif(i==5):
        print('Guitar_electronic')
    elif(i==6):
        print('keyboard_acoustic')
    elif(i==7):
        print(' keyboard_elec')
    elif(i==8):
        print(' keyboard_synth')
    elif(i==9):
        print(' mallet')
    elif(i==10):
        print(' organ')
    elif(i==11):
        print(' reed')
    elif(i==12):
        print('string')
    elif(i==13):
        print('vocal_synth')
```

```
The instruments given for predicitons are:
Flute
```

The samples given for Guitar_Acoustic are:




The output for the samples is as shown below.

```
print('The instruments given for predicitons are:')
for i in Label_Data_Test[2:3]:#predictions:
    if(i==0):
        print('Bass_Acoustic')
    elif(i==1):
        print('Bass_electronic')
    elif(i==2):
        print('Bass_Synthetic')
    elif(i==3):
        print('Flute')
    elif(i==4):
        print('Guitar_Acoustic')
    elif(i==5):
        print('Guitar_electronic')
    elif(i==6):
        print('keyboard_acoustic')
    elif(i==7):
        print(' keyboard_elec')
    elif(i==8):
        print(' keyboard_synth')
    elif(i==9):
        print(' mallet')
    elif(i==10):
        print(' organ')
    elif(i==11):
        print(' reed')
    elif(i==12):
        print('string')
    elif(i==13):
        print('vocal_synth')
```

```
The instruments given for predicitons are:
Guitar_Acoustic
```

The above results are obtained because of testing the datasets. In the same way, the results can be shown for all the music instruments we've used considering the samples. The check is made to obtain the correct prediction or identification of the instrument. The confusion matrices obtained from ANN and CNN analysis are because of the training of the dataset which determines the accuracy and the prediction result shows the correctness of the recognition of instrument.

## VI. CONCLUSIONS

The confusion matrices obtained from the ANN and CNN analysis are compared. It is found that CNN is more efficient compared to ANN as the accuracy is more in a Convolution Neural Network. CNN is more efficient in both memory and complexity. Imagining a billion neurons in a normal artificial neural network would eat up a lot of memory and would be very highly complicated with the layers being present, whereas Convolutional Neural network reduces the memory consumed and also the complexity of the neural network is reduced. Also, Convolutional Neural networks outperform the ANN in image and audio recognition. However, it takes a lot of time to train and test a data in CNN as compared to ANN. It shows the time complexity is more in CNN. But, the other merits of CNN clearly overpower this disadvantage of CNN. It is no wrong to say that this music instrument recognition model can further be used in music genre classification, tone identification and also pitch identification.

## REFERENCES

[1] DimitrosGiannoulis, EmmanoulBenetos, AnssiKlapuri and Mark D Plumbley, "Improving Instrument recognition in polyphonic music through system integration", IEEE International Conference on Acousics, speech and signal processing (ICASSP), 2014.

[2] Soren Bekher, Marcel Ackermann, Sebastian Lapuschkin, Klaus Robert Muller, "Interpreting and explaining deep neural networks for classification of audio signals".

[3] Aditya Khampariya, Deepak Gupta, Ashish Khanna, Babita Pandey, PrayagTiwary, "Sound Classification using CNN and Tensor deep stacking network", *Institute of Electrical and Electronics Engineers (IEEE),* India 2018.

[4] Jongpil Lee, Taejun Kim, Jiyoungpark, Juhannam, "Raw waveform based audio classification using sample level CNN architectures", 2017.

[5] Chandan S V, Mohan R Naik,andKarthik D P, "Indian Music Instrument Identifier for polyphonic audio signal", 2018.

[6] Zhouyu Fu, Guojun Lu, Kai Ming Ting and DengshengJhang, "A survey of Audio-based music classification and annotation", China,2001

[7] TheodrusTheodrou, IosifMporas and Nikos Fakotakis, "An overview of automatic audio segmentation", Greece, 2014

[8] Monali R. Pimpale, Shanthi Therese and Vinayak Shinde, " A survey on: Sound Source Separation Methods", Mumbai, 2016.

[9] Jacob O'Bryant, "A survey of music recommendation and possible improvements", 2017.

[10] Yoonchang Han, Jaehun Kim and Kyogu Lee, "Deep Convolutional Neural Network for Instruction recognition in polyphonic music", 2017.

[11] SiddharthSigtia, EmmanouilBenetos and Simon dixon, ,"An end to end neural network for polyphonic music transcription", 2016.
[12] ArieLivshin and Xavier Rodet, "Purging Music instrument sample databases using automatic music instrument recognition methods", 2019.