International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Comparing the Accuracy of CNN Model with Inception V3 for Music Instrument Recognition

Renju K.[1], Ashok Immanuel V.[2]

Submitted: 24/12/2023 **Revised**: 30/01/2024 **Accepted**: 06/02/2024

Abstract: Identification of music instruments from an audio signal is a complex but useful task in music information retrieval. Deep Learning and traditional machine learning models are extremely very useful in many music related tasks such as music genre classification, recognizing music similarity, identifying the singer etc. Music Instrument recognition and classification would be helpful in categorizing different categories of music. Many researchers have proposed models for classifying western music instruments. But very little research has been done in identifying instruments accompanied with South Indian music. This research aims at identifying string instrument such as violin and woodwind instrument such as flute accompanied in a Carnatic music concert and also in other categories of music. In order to identify the instruments accompanied, Convolutional Neural Network model and Inception V3 models were used. The Mel Frequency Cepstral Coefficients images were extracted from the audio input and fed in to the neural network model. The model has been trained for the above mentioned instruments, tested and validated on different types of audio input. This research also evaluates the performance of Inception V3 transfer learning model with CNN model in recognizing the instruments used in different categories of music.

Keywords: *Music Instrument recognition, Mel Frequency Cepstral Coefficients, Machine Learning, Deep Learning*

1. Introduction

Musical Instrument recognition from a polyphonic audio signal is a complex task in audio signal processing. Automatic identification of music instruments is a solution to many music related tasks such as music genre classification, instrument classification, vocal source separation, assessing the harmony of vocals and instruments in music information retrieval. An important aspect of instrument identification is in discovering the exact features of sound which would identify the instrument. It remains a challenge in music signal processing. The main difficulty in instrument recognition and classification from polyphonic sounds is that it contains a mixture of different sounds and these sounds interfere with each other. Hence It would be difficult to extract the acoustic features of instruments efficiently and accurately. Music Instrument sound is usually characterized by properties such as pitch, loudness and timbre [22,25]. Though pitch and loudness would be the same for two different instruments, timbre property would be distinct for instrument playing with same pitch and The features such as Spectrogram, Mel Frequency Cepstral Coefficients(MFCC), Spectral Centroid, Linear Predictive Coding were used by many researchers in recognizing instruments from an audio signal [1]. Out of these features, MFCC is widely used feature as it is closely related to the human perception of sound [3,5,13,22]. Researchers have proved that MFCC plays a vital role in discovering the formants and timbre of sound [17]. The rate of change of spectral band of an audio signal is given by its cepstrum. This conveys different values that construct the timbre and formant of audio signal. The twenty MFCC coefficients extracted from an audio input gives the most relevant information regarding formants, spectral envelope and timbre of sound which would be unique for each instrument. The process of extracting MFCC from an audio input [4,9] is given in Fig 1.

[1] *Department of Computer Science,Research Scholar, CHRIST(Deemed to be University), Karnataka, India.*
[1] *Department of Computer Science, Assistant Professor, Mount Carmel College, Autonomous, Karnataka, India.*
[2]*Department of Computer Science, Associate Professor, CHRIST(Deemed to be University), Karnataka, India*
[1]*renju.k@res.christuniversity.in*

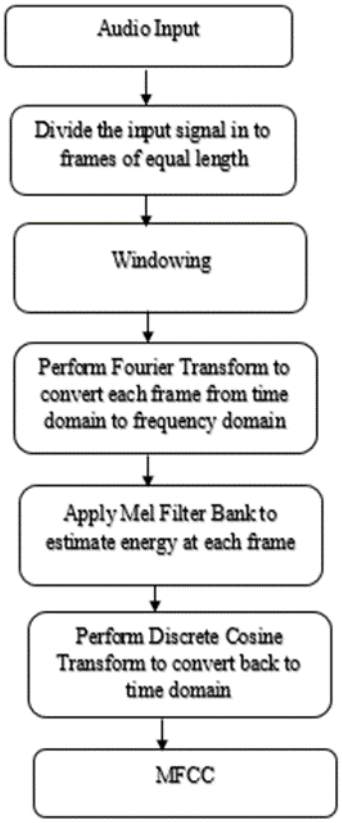

Fig 1: Extraction of MFCC from Audio Signal

2. Related Work

Many researchers have proposed deep learning models models for instrument classification [10,11,12,14,18]. Christine Dewi et.al used YOLO V4 for object detection in identifying music instruments by drawing a bounding box for all instruments using BBox mark tool [2]. The YOLO V4 algorithm separates the image in to NXN grids with each grid capable of generating K bounding boxes. It would compute B boundary boxes for every grid cell along with a confidence score. The Convolution Neural Network model was used to predict the class by estimating the bounding boxes which contains the image of the instrument. They have used PPMI dataset which contains the photographs of people using instruments. The images of instruments such as Bassoon, cello, clarinet, French horn, erhu, flute, guitar, harp, saxophone, trumpet, recorder, and violin were incorporated in the dataset. Apart from YOLO V4, Densenet was also used for object detection and results were compared They have chosen instruments which looked similar for the research, yet got a maximum average accuracy of 94.7% with CNN model.

Saranga et.al proposed a model for instrument recognition only using one feature which is Mel Frequency Cepstral Coefficients of audio data [3]. They have developed an artificial neural network model which would classify twenty different classes of instruments belonging to different categories such as woodwind, brass, percussion and string instruments. As the dataset used was London Philharmonic instrument dataset which was highly imbalanced, they have implemented stratified splitting which would split the data proportionally in to training and validation datasets. The model showed an accuracy of 99% for training and 97% for validation data. In [4], the authors have explained in detail the audio pre processing steps such as removal of silence from audio, pre-emphasis, segmentation, framing, windowing, feature analysis and recognition in detail. They have considered many features of audio signal for instrument recognition such as fundamental frequency, MFCC, Chroma, Spectral Skewness, Spectral Spread, Temporal Centroid, Zero Crossing Rate, Spectral Roll Off, Crest Factor Spectral Flux and Spectral Centroid. The ANN classifier was used to classify six different types of instruments and got an accuracy of 95% in their result.

The authors of [6] have developed a conventional machine learning K Nearest Neighbors(KNN) model for music instrument sound classification. The MFCC feature of audio were extracted from databases with distinct characteristics for classification of instruments. The instruments modelled for classification were violin, Flute, Trumpet and Guitar. The KNN model discovered an accuracy of 92% which proved that traditional machine learning models are also better in classifying instruments. Romulo et.al evaluated the performance of classifying instruments with Naïve Bayes, Decision Tree and Support Vector machine supervised learning algorithms [7]. They have observed 92% accuracy for decision tree model while Naïve Bayes and Support Vector Machine models gave an accuracy of 69% and 75% respectively. To improve the accuracy for SVM and Naïve Bayes models, more number of features were incorporated but the accuracy dropped up to 25%. This proved that increasing the number of features would not always improve the accuracy of the model.

Music emotion recognition model was developed by Sangeetha et.al, extracted MFCC, spectral features such as centroid, bandwidth, roll off. and temporal feature such as zero crossing rate of audio signal [8] for emotion recognition. They have classified the instruments such as string, percussion, woodwind and brass in to four emotions such as happy, sad, neutral and fear respectively. The Recurrent Neural Network(RNN) model was developed and compared it with other machine learning models for the features extracted. They discovered that MFCC with deep RNN gave better performance in music instrument emotion recognition. In [9], authors have extracted 39 MFCC features, 22 Sonogram and 61 MFCC combined with Sonogram features were extracted from the audio signal. They have evaluated the performance of SVM with MFCC., SVM with Sonogram, SVM with MFCC and Sonogram, KNN with MFCC, KNN with Sonogram and KNN with MFCC and Sonogram classifier models. They found that SVM with MFCC and Sonogram gave a higher accuracy of

98% when compared to all other classifier models mentioned above.

Researchers have also evaluated the performance of Multilayer Perceptrons (MLP) Convolutional Neural Network(CNN) and Convolutional Recurrent Neural Network(CRNN) models for instrument activity detection from polyphonic signals [11]. They could classify 18 different instruments and proved that CNN and CRNN models outperform MLP models in music instrument classification. The audio features such as Enhanced Mel Frequency Cepstral Coefficients and Enhanced Power Normalized Cepstral Coefficients were extracted in order to develop a classifier model given in [16]. They have compared the performance of different classifier models such as J48, K Star, Random Forest, Bagging and BFTree and got a higher accuracy of 98% for Random Forest model. The least accuracy was observed for BFTree model in instrument classification.

In [22], authors have extracted features such as MFCC along with Spectral Centroid, Spectral Roll off, Energy, Linear Prediction Coefficients, Spectral Flux features for music instrument recognition. They have categorized the instruments in to string, woodwind, brass and keyboard types which includes 16 different types of instruments. The dataset had 16 .wav files of different instruments with ten C, D and E notes in particular. They have proposed a new model based on the wavelet packet transform by computing the wavelet coefficients at different frequency sub bands with different resolutions. After pre-processing of signal, the wavelet packet transform feature is extracted from the audio. It used 16 bands with 4 level of decompositions and then computed the energy at each frequency band to the energy at the first frequency band for normalizing the parameters. Then the statistical parameters such as mean, standard deviation and variance were calculated for the classification of instruments.

3. About the Dataset

The Instrument dataset used for this research was manually generated, taken from various open source repositories such as myfreemp3, mobcup and also from youtube videos. The real challenge was in the generation of dataset as it was not easily available according to our requirement. We had altogether 300 audio files with each instrument having 100 files approximately for training and validation. The instruments used in Carnatic music concert and also in various categories of South Indian music such as violin and flute were used for this research. This data was used for training and validation and the data used for testing included the vocals accompanied with different instruments and only instruments without any vocals as well. The testing data was generated manually from various open source repositories which contains a mixture of Carnatic and film songs. The training and testing data were divided in the ratio of 80:20.

4. Flow of the Work

The Python library Librosa was used to load the audio file. All audio files were in mp3 format and these files were converted in to .wav format for processing. As the audio signal is continuous and varying, we need to divide the signals in to frames. From each frame, audio samples were taken for analysis. The rate at which the signals were sampled was 22050Hz and the duration of audio file considered was 60s. The Python library Librosa has extensive set of methods which were used for audio pre-processing. As the dataset contains polyphonic sounds, a fast and reliable music source separation tool spleeter [15] was used to extract the sounds. This extraction is necessary because the polyphonic sounds contains a mixture of many instrument sounds. Also, these sounds interfere with each other and makes it difficult for the model to understand. Keeping in mind the speed, clarity and ease of use, the spleeter model was used. The spleeter model divides the audio signal in to 2 stems for vocal/accompaniment separation, 4 stems for vocals, drums, bass and other separation and 5 stems for vocals, drums, bass, piano and other separation[15] Fig 2. For this research, we have used 4 stems separation which divided the audio signal in to vocals, drums, bass and other.

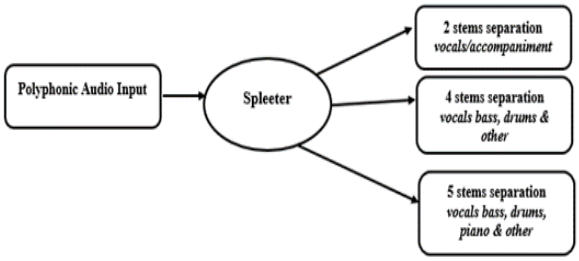

Fig 2: Vocal Source Separation Using Spleeter

The audio files were already noise free and not much of complex preprocessing steps were implemented. To train the model for Instrument recognition, the MFCC images having 20 coefficients of each audio instrument file from the dataset were extracted and saved it in the directory named as Yes/No according to the recognition of the respective instrument. The 'Yes' directory contained the MFCC images of files which had the respective Instrument played while the 'No' directory contained the images where the respective instrument were not played. There were a total of 150 MFCC images in both directories. These images were fed in to the neural network and the model was developed which could recognize the presence of instruments such as violin and flute The process flow is given in Fig 3.

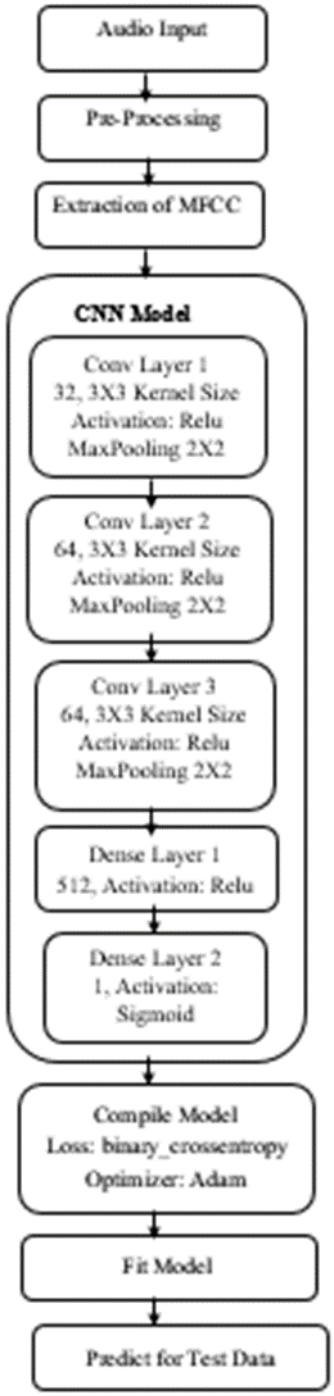

Fig 3: Recognition of Musical Instruments Using Convolutional

Neural Network Model

A simple Convolutional Neural Network model consisting of 3 layers, with the first layer having 32 filters, 3X3 kernel size, second and third layer having 64 filters, 3X3 kernel size were used for instrument recognition. A maxpooling function was applied at the end of each convolutional layer to retrieve the maximum value over an input window and the same is passed to the next layer. The activation function as shown in Fig 5. The model was then evaluated for the test data and witnessed an accuracy of 74%. The precision and recall metrics were 100% and 25% respectively for the recognition of flute instrument.

ReLu was used in every convolutional layer which would retain the positive values and assign zero to all negative values, if any. The two dense layers, one had 512 neurons and the output dense layer had 1 neuron with activation function sigmoid, for recognizing whether the instrument is present or not in the audio file. The model was compiled with binarycrossentropy as loss function and optimizer used was adam. The model's accuracy was measured for training and validation data and then predicted for the test data.

As we know the challenge with deep learning model is that it requires huge volume of data for training in order to get good accuracy. Even if we have got a large dataset for training, it may take many days or even weeks to complete the task [26]. To overcome this problem, researchers have discovered pretrained models such as Inception V3, ResNet, DenseNet which were already trained on ImageNet dataset contained millions of data and the weights of these models could be reused to solve the current problem. This concept called transfer learning where a new model is developed by taking the advantage of pretrained models in order to reduce the training time and for getting better accuracy [18]. We have used Inception V3 pretrained model for the recognition of musical instruments and compared the results obtained with CNN model. The Inception V3, consists of 42 layers is a modified version of Inception V1 and V2 [26] were used to train on the instrument dataset. The input and dense layers of the pretrained model were customized with respect to the current problem.

5. Results and Discussion

The test dataset contains the MFCC images of different types of music, Carnatic and film songs. The developed models, one with Inception V3 transfer learning model and the other model with CNN were tested for the test data. The flow_from_ditrectory() method of ImageDataGenerator class was used to iterate through the images from the training, validation and test datasets. All the images were scaled to 200X200 size, set the batch size as 10 and the class mode as binary.

5.1) CNN Model

With CNN model, the accuracy ,of training and validation data were only 72% and 74% respectively for the recognition of string instrument violin as shown in Fig 4. The model was tested for the test data which contained MFCC images of audio and observed an accuracy of 70%. The metrics precision and recall were 81% and 75% respectively for the detection of violin instrument. The training and validation accuracy of detecting flute instrument was not consistent and after 18 epochs, the model gave a consistent accuracy of 80%

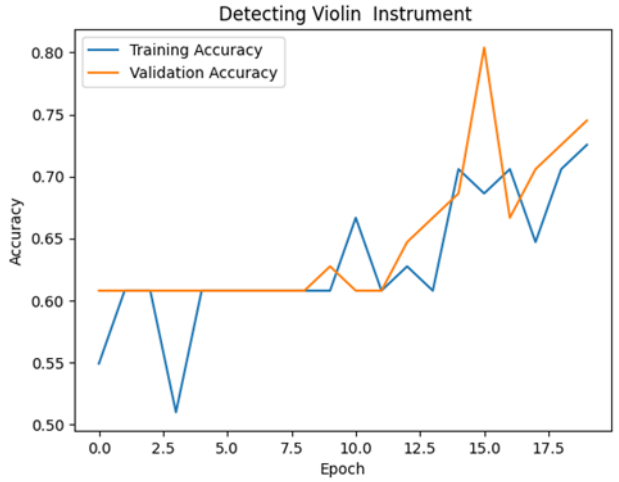

Fig 4: Accuracy of Training and Validation data for the recognition of Violin Instrument

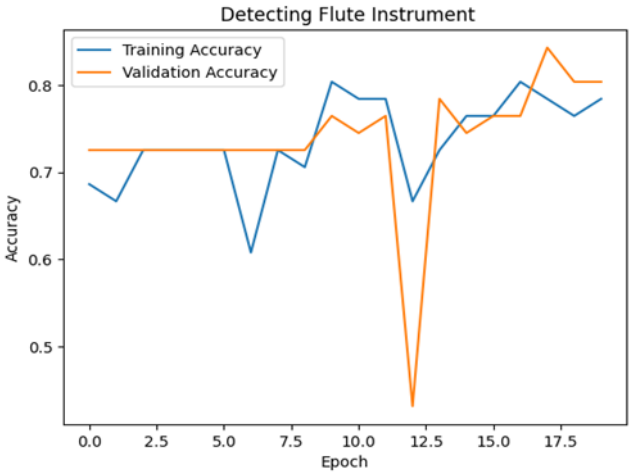

Fig 5: Accuracy of Training and Validation data for the recognition of Flute Instrument

5.2) InceptionV3 Transfer Learning Model

With the intention to improve the accuracy of CNN model, we have also implemented transfer learning model with Inception V3 for instrument recognition from polyphonic audio input. The pretrained Inception V3 model gave an accuracy of 100% for the training and validation data for the recognition of instruments violin and flute as shown in Fig 6 and Fig 7. The Inception V3 model was then evaluated for the test data which gave an accuracy of 87% for the recognition of violin instrument. The precision and recall metrics were 80% and 100% respectively. For the recognition of flute instrument, the Inception V3 model gave an accuracy of 92% and the precision and recall metrics were 100% and 78% respectively. The results are summarized in Table 1.

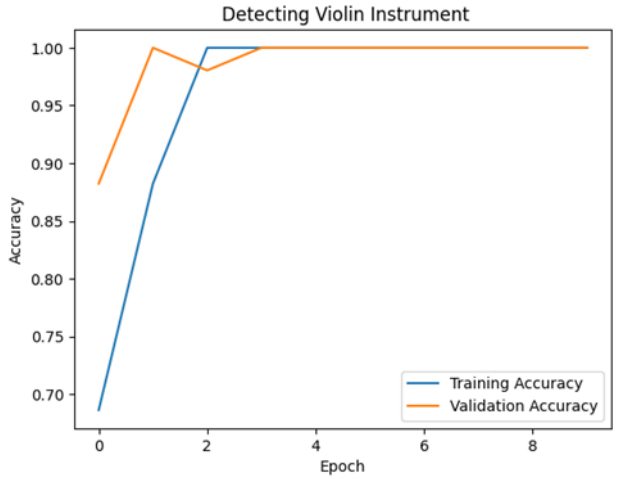

Fig 6: Accuracy of Training and Validation data for the recognition of Violin Instrument with Inception V3

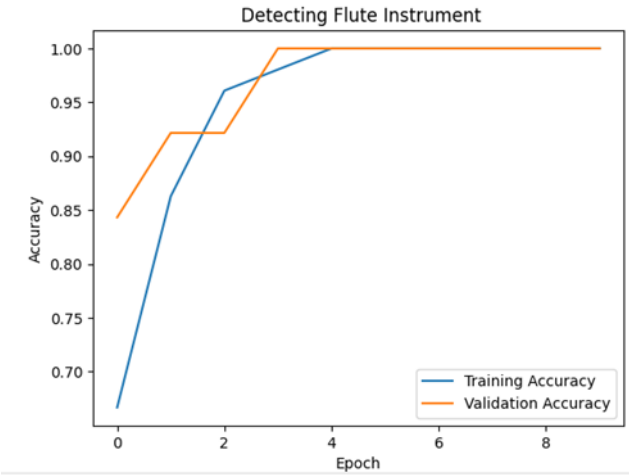

Fig 7: Accuracy of Training and Validation data for the recognition of Flute Instrument with Inception V3

Table 1: Evaluation of Metrics in Recognizing Instruments Using CNN and Inception V3 Models

Recognition of Instrument	**Metrics**	**CNN Model**	**Inception V3 Model**
Violin	Accuracy	70%	87%
	Precision	81%	80%
	Recall	75%	100%
Flute	Accuracy	74%	92%
	Precision	100%	100%
	Recall	25%	78%

6. Conclusion

Music instrument recognition has been a hot research topic in music information retrieval. Many researchers have proposed models for music instrument recognition. This novelty of this research lies in the fact that with minimal pre-processing techniques, music instruments such as violin and flute were identified and recognized from different categories of music using deep learning techniques. The categories of music considered were Carnatic music accompanied with different instruments, film songs and songs only with instruments. This research also evaluated the performance of CNN model with the pretrained model Inception V3 and found that Inception V3 model outperformed CNN model with a better accuracy The future enhancement would be to track the pitch of an instrument from a polyphonic audio signal.

References

[1] K. Sreekar, A. Devansh Reddy, "Musical Tones Classification using Machine Learning", *International Journal for Research in Applied Science & Engineering Technology*, ISSN: 2321-9653. Volume 10, Issue XII, December 2022.

[2] Christine Dewi , Rung-Ching Chen, "Deep Learning for Advanced Similar Musical Instrument Detection and Recognition", *International Journal of Computer Science*, Volume 49, Issue 3, September 2022.

[3] Saranga Kingkor Mahanta , Abdullah Faiz Ur Rahman Khilji , Partha Pakray, "Deep Neural Network for Musical Instrument Recognition using MFCCs", *Research Gate, http://dx.doi.org/10.13053/CyS-25-2-3946*

[4] Sujith. S, Jithin Kumar I J, Appu.V. "Musical Instrument Recognition System with Dynamic Feature Selection Method", *International Journal of Engineering Research & Technology*. ISSN: 2278-0181. Volume 10, Issue 05, May-2021.

[5] P. Aurchana, S. Prabavathy, "Musical Instruments Sound Classification using GMM", *London Journal of Social Sciences*. Volume 1, 2021.

[6] P. Kathirvel, R. Thiruvengatanadhan, R. Seenu, "Music Instrument Sound Classification using KNN*", International Journal of Creative Research Thoughts,*ISSN: 2320-2882. Volume 9, Issue 12, December 2021.

[7] Romulo Vieira , Joao Teixeira Ara ujo , Edimilson Batista , Flavio Schiavoni , "Automatic classification of instruments from supervised methods of machine learning", *SBC-OpenLib*, October 2021.

[8] Sangeetha Rajesh,N J Nalini, "Musical Instrument Emotion Recognition Using Deep Recurrent Neural Network", *Elsevier*, Volume 167,2020.

[9] S. Prabavathy, V. Rathikarani, P. Dhanalakshmi, "Classification of Musical Instruments Using SVM and KNN", *International Journal of Innovative Technology and Exploring Engineering*., ISSN:2278-3075. Volume 9, Issue 7, May 2020.

[10] Alexandre M. Lucena, Caroline P. A. Moraes, Kenji Nose-Filho , Denis G. Fantinato, "Musical Instruments Recognition using Machine Learning Techniques MLP and SVM", *Brazilian Technology Symposium*, October 2020.

[11] Siddharth Gururani, Cameron Summers, Alexander Lerch,. "Instrument Activity Detection in Polyphonic Music Using Deep Neural Networks", *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR*), At: Paris, France. April 2019.

[12] Arun Solanki, Sachin Pandey, "Music Instrument Recognition Using Deep Convolutional Neural Networks", *International Journal of Information Technology, Springer*. January 2019.

[13] Orchisama Das, "Musical Instrument Identification with Supervised Learning", *Semantic Scholar, Corpus ID: 203603297*. 2019.

[14] S. Prabavathy, V. Rathikarani, P. Dhanalakshmi, "An Enhanced Musical Instrument Classification using Deep Convolutional Neural Network", *International Journal of Recent Technology and Engineering*, ISSN:2277-3878.Volume 8,Issue 4,November 2019.

[15] Romain Hennequin, Anis Khlif, Felix Voituret, Manuel Moussallam, "Spleeter: A Fast and State-Of-The Art Music Source Separation Tool with Pre-Trained Models", *ISMIR* 2019.

[16] A.Muthumari, "Classification Analysis for Musical Instrument Signal", *International Conference for Phoenixes on Emerging Current Trends in Engineering and Management*, 2018.

[17] Shruti Chakraborty, Ranjan Parekh, "Improved Musical Instrument Classification Using Cepstral Coefficients and Neural Networks", *Semantic Scholar. DOI:10.1007/978-981-13-2345-4_10 Corpus ID: 70343606*, 2018.

[18] Long Nguyen,Dongyun Lin, Zhiping Lin, "Deep CNNs for Microscopic Iimage Classification by Exploiting Transfer learning and Feature Concatenation", *IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2018

[19] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous,

Bryan Seybold, Malcolm Slaney, Ron J. Weiss, Kevin Wilson, "CNN Architectures For Large-Scale Audio Classification", *Semantic Scholar*, DOI:10.1109/ ICASSP.2017.7952132 , Corpus ID: 8810481. January 2017.

[20] Sumit S. Bhojane, Omkar G. Labhshetwar, Kshitiz Anand, Prof. S. R. Gulhane, "Musical Instrument Recognition Using Machine Learning Technique", *International Research Journal of Engineering and Technology*, Volume 4, Issue 5, May 2017.

[21] Taejin Park , Taejin Lee, "Musical Instrument Sound Classification With Deep Convolutional Neural Network Using Feature Fusion Approach",. *arXiv:1512.07370*, December 2015.

[22] Abhilasha, Preety Goswami, Prof. Makarand Velankar. "Study paper for Timbre identification in Sound", *International Journal of Engineering Research & Technology*, ISSN: 2278-0181. Volume 2 Issue 10, October – 2013

[23] Dr. D.S. Bormane, Ms. Meenakshi Dusane. "A Novel Techniques for Classification of Musical Instruments", *Information and Knowledge Management*. ISSN 2224-5758. Volume 3, No.10, 2013.

[24] Stephen McAdams, "Musical Timbre Perception The Psychology of Music", *Elsevier*, December 2013.

[25] Glenn Eric Hall , Hassan Ezzaidi , Mohammed Bahoura, "Study of Feature Categories for Musical Instrument Recognition", *Communications in Computer and Information Science Book Series*. Volume 322,2012.

[26] https://machinelearningmastery.com/how-to-use-transfer-learning-when-developing-convolutional-neural- network-models/