Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Performance Evaluation of a CNN-Based Deep Learning Model for Classifying String Musical Instruments Using Sound Data and Various Feature Extraction Techniques

Shital Santosh Pawar[1], Dr. Narendra Sharma[2]

[1]*Research Scholar, Sri Satya Sai University of Technology and Medical Sciences, Sehore, MP*
[2]*Research Guide, Associate Professor, Department of CSE, Sri Satya Sai University of Technology and Medical Sciences, Sehore, MP*
Email: svghotekar@gmail.com

Musical Instrument Recognition is a subfield of machine learning and signal processing that focuses on automatically identifying musical instruments from audio recordings. Music combines sound generated by multiple instruments in terms of plenty of audio signals. The main purpose of this study is to construct an extensive framework for audio signal analysis, property extraction, behaviour prediction, and pattern recognition concerning the sounds generated by musical instruments. This paper intends to explore the machine and deep learning algorithms to implement the constructed model. It involves extracting meaningful features from the audio signal and using them to classify the instrument. This article describes a convolutional neural networks (CNNs) approach for classifying 7-string musical instruments to learn complex patterns directly from the raw audio signal. This paper addresses the challenges of recognizing multiple musical instruments simultaneously using traditional feature extraction methods such as Mel-Frequency Cepstral Coefficients (MFCCs) and spectrograms, and deep learning-based approaches such as CNN. Sound classification of multiclass instruments is performed under the classes like banjo, cello, double bass, guitar, mandolin, viola, and violin. The overall performance of the CNN model is calculated in terms of Accuracy, Macro Avg F1 Score and Weighted Avg F1 Score is 98% respectively.

Keywords: Deep Learning, Feature Extraction, String Musical Instruments, MIR, Audio Signal Processing.

1. Introduction

In this digital era music plays an important role in entertaining people, reducing stress levels, providing peace and harmony, relaxing, and improving mood. It is also used in music therapy for patients. Learning to play an instrument or listen to music, especially in a young age group demonstrated to improve memory, focus, and problem-solving abilities. It also helps people maintain rhythm and energy during workouts or athletic performances. Music has a significant impact on lifestyle, fashion, and artistic trends. Music may also play a part in social events, religious rites, cultural activities, and festivals, among other things. Music is deeply integrated into many aspects of human life like providing emotional, cognitive, social, and physical benefits. Some common applications of musical information retrieval are instrument identification, genre classification, playlist generation, music recommendation systems, and music transcription. Music is made up of various instruments. Different kinds of instruments are present in Indian musical cultures, such as String, Woodwind, Percussion, and Brass. All these instruments can be used separately or combined to form music, songs, and albums. Each type of instrument has its unique sound which can be understood by using various properties like pitch, texture, duration, loudness, and timbre [1].

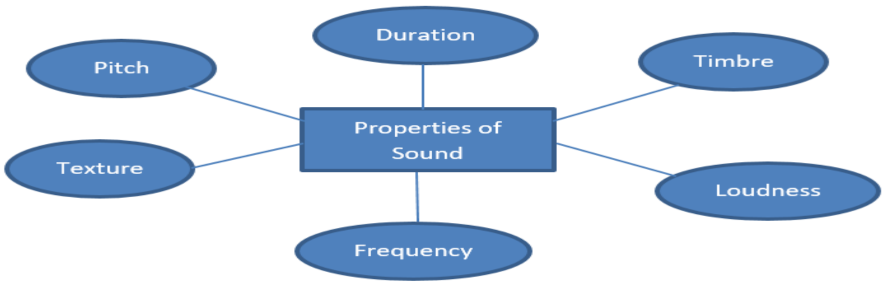


Figure:1 Properties of sound

Pitch- It is the highness or lowness of a sound.

Texture- It's related to the combination of different frequencies and their amplitudes over time.

Duration- Means how long a sound continues.

Loudness- It is the intensity or volume of a sound.

Timbre- It is a unique quality that distinguishes one instrument from another, even when playing the same note at the same pitch.

Musical instruments are from different families like woodwind with string or brass with percussion or string with bass etc. then it is easily identified as compared to instruments are with the same family. Audio information extraction is a challenging task, in both monophonic and polyphonic music so to overcome this problem our research is focused on musical instrument identification from the string family. String musical instruments are the oldest and usually used in the largest section of an orchestra and consist of violins, violas, cellos and double basses, etc. It is difficult to extract features, waveforms, frequency, pitch, and timbre, especially from polyphonic music that includes instruments from the same family [1]. Automatic classification of musical instruments using acoustic features two modeling techniques SVM & KNN are used. It uses MFCC and Sonogram with SVM to identify the musical instruments with 98% of accuracy on sixteen musical instruments [2]. The sound generated by monophonic music is identified in this study using spatiotemporal features and comparing the NN using cepstral features. The performance of bidirectional Long Short Term Memory (Bi-LSTM) network,1-D CNN, and 2-D CNN models obtained a macro F1 score of

0.976,0.977 and 0.985 respectively [3]. The study evaluates features for musical instrument recognition from monophonic musical signals Quadratic Discriminant Analysis shows the lowest error rate (7.19% for individual instruments and 3.13% for instrument families) over a dataset of 1007 tones from 27 musical instruments without employing any hierarchical structure [4]. The study proposes a novel method for automatically identifying all instruments in an audio excerpt using individual convolutional neural networks. The model efficiency is high, with metric values ranging from 0.86 for the guitar to 0.99 for the drums [5]. The authors found that SVMs and QDA performed the best, achieving success rates close to 70% in classifying 27 different instruments. Brass and woodwind instruments were generally easier to classify than string instruments [6].

This paper is aimed to throw light on the feature extraction techniques for sound classification as well as machine and deep learning terminologies. The authors have structured a proposed framework for musical instrument classification. This research paper is structured with multiple sections as follows. Section II provides a review of work regarding feature extraction and classification of musical instruments using machine and deep learning. Section III describes the machine learning and deep learning paradigm. Section IV outlines some feature extraction approaches for sounds from musical instruments. The proposed architecture for the classification of musical instrument sound designed by the author is given in Section V. The results of the experiment are described in Section VI. In the end, Section VII concludes the research work carried out here and VII projects the future enhancement.

2. Related Work

In this section, we have gone through the work done previously in this context of music identification with respect to Machine Learning and Deep Learning Methods.

Blaszke and Kostek et al. [7] propose a novel approach for automatically identifying single instruments within a musical composition. Their method utilizes a set of Convolutional Neural Networks (CNNs), each trained to recognize a specific instrument. A separate CNN is trained for each instrument, allowing for more specialized feature extraction and recognition. The proposed method achieves high accuracy in identifying various instruments, with metrics like precision and recall ranging from 0.86 to 0.99.

Solanki and Pandey et al. [8] propose a deep convolutional neural network (DCNN) architecture for musical instrument recognition. Their approach focuses on recognizing the predominant instrument in polyphonic music. The authors use a DCNN with multiple layers to extract features from the audio signal. The audio signal is converted into a Mel spectrogram, which represents the spectral power distribution of the signal. The DCNN automatically learns relevant features from the Mel spectrogram, such as timbre and spectral characteristics. The final layer of the DCNN is a softmax layer, which predicts the probability of each instrument being the predominant one.

Joshi, Pareek, and Ambatkar et al. [9] conducted a comparative study to evaluate the performance of Mel-Frequency Cepstral Coefficients (MFCCs) and Mel spectrograms for raga classification using a Convolutional Neural Network (CNN). Both MFCCs and Mel spectrograms are effective features for raga classification. The authors employed a CNN architecture with multiple convolutional layers and pooling layers to extract relevant features from the input data (MFCCs or Mel spectrograms).

Monica S. Nagawade, Varsha R. Ratnaparkhe et al. [10] highlight the MFCCs are extracted from the audio signal to represent the spectral envelope of the sound. A Support Vector Machine (SVM) classifier is used to distinguish between different musical instruments based on the extracted MFCC features. There are 90 samples altogether (60% of all samples) in the training database, while 60 samples (40%) are utilized in the testing database. The K-NN classifier was employed during the classification phase. The accuracy of the proposed method is 91.66% for the cello, piano, and trumpet and 83.33% for the violin and flute.

Sarfaraz Masood, et al. [11] The purpose of this study is to address the issue of detecting musical instruments in monophonic audio samples with a novel feature extraction method based on the Short-Time Fourier Transform (STFT). A multilayer perceptron (MLP) neural network is used to classify musical instruments based on the extracted features. The proposed method gives an accuracy of 92% in identifying various musical instruments.

Seyed Muhammad Hossein Mousavi et al. [12] proposed a method for recognizing Persian classical musical instruments using a newly created dataset comprising 7 classes viz Santur, Kamancheh, Ney, Tar, Tonbak, Ud, and Setar. Features like Spectral Roll-off, Spectral Centroid, Zero Crossing Rate, Entropy Energy, and Mel-Frequency Cepstrum Coefficients (MFCCs) were extracted from an audio signal. The Multi-Layer Neural Network (MLNN) performs the classification task, while the Fuzzy Entropy Measure is used in the feature selection portion. Fuzzy entropy was used to select the most informative features. The proposed method achieved an accuracy of 92.7% in recognizing Persian classical musical instruments.

D. M. Chandwadkar, M. S. Sutaone et al. [13] This study discusses the importance of several characteristics with different classifiers in automatically identifying musical instruments. The authors experimented with various feature extraction techniques, including Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coding (LPC), and Zero Crossing Rate (ZCR) to identify the violin, xylophone, piano, and acoustic guitar. Different classifiers were evaluated, such as Support Vector Machines (SVMs), Naive Bayes, and K-nearest neighbours (KNN). The best combination of features and classifiers gives 90% of accuracy.

Kumar et al. [14] proposed a hybrid approach combining classical machine learning (ML) and deep learning (DL) techniques to identify morbidities from clinical notes using both traditional ML features (e.g., TF-IDF, word embedding) and DL features (e.g., word2vec, BERT) with 82.9% of accuracy.

Norhalina Senan et al. [15] the paper demonstrate the effectiveness of MFCCs and SVMs for musical instrument identification, even in the context of traditional Malay music. This study used the perception-based and MFCC features schemes, which had a total of 37 characteristics. The changed data sets and derived feature schemes are assessed for classification performance using a Multi-Layered Perceptron's classifier. Results indicate that the data sets with the combination of complete characteristics produced the highest accuracy, 99.57%.

3. Machine Learning and Deep Learning Paradigm

Machine Learning and Deep Learning algorithms have dominantly contributed to the field of sound classification. In traditional programming, the system accepts input and generates output based on the logic. In Machine Learning, the system accepts both input and output and generates models. That model is used to produce predictions and solve complex challenges such as data analysis, business problems, and real-world concerns. Classification in machine learning is based on previously made observations and pre-labelled training data. Deep learning is one of the most well-known applications of machine learning, which is made up of a variety of algorithms and methods that are helpful for complicated data (Text, Voice, Image) classification. High classification accuracy might potentially be provided by deep learning systems. [16]

Speech processing has undergone a revolution because of deep learning architectures, which have shown outstanding performance on a variety of tasks. Deep learning models have outperformed conventional methods in fields like speech recognition because of their capacity to autonomously learn hierarchical representations from unprocessed speech data.

Convolutional Neural Network

CNN is one of the classifier of Deep Learning algorithms and which is widely used for classification of various kinds of data. Convolutional Neural Networks, like neural networks, adjust their weights over time to facilitate learning. CNNs are multi-layer networks that incorporate hidden layers including convolution, pooling, and fully linked layers, as well as input and output layers. The convolutional neural network mimics the organisation of human neural tissue, including the convolution layer and accompanying subsampling layers. The system establishes a forced and local association between upper and lower layers, as well as nearby brain tissues, using certain regulations. These units provide a continuous receiving domain in space. Convolutional neural networks have the ability to reduce network parameters and learn local features [17].

4. Feature Extraction Techniques used for identification of musical instruments using sound

Feature extraction is the process of converting input data into a set of condensed digital representations of feature schemes. The goal of feature extraction is to extract pertinent information from the input data so that a specific job may be carried out using a small collection of features rather than the entire set of huge data. Previous research has used numerous feature approaches to classify the sound of musical instruments. It is done either individually or in combination. Two types of characteristics are employed which include spectral and temporal information [18].

Time-domain and spectral-domain features are used for the classification of musical instruments.

Time-domain features (Temporal Features) includes features related to the shape of the envelope of a note. Time-domain features are measured as the attack time(A), decay time(D), sustain time(S), and release time (R).

Attack time(A) is a time difference between the onset time and end-of-attack. Onset time is the time at which the note begins to sound. End-of-attack is the time when the envelope reaches maximum value.

Decay time(D) is the time difference between end-of-attack and the forward position where the Amplitude is 25% of the Amplitude at end-of-attack.

Spectral Features (Timbral Features) include the quality of a musical sound which is used for differentiating two sounds when they are in the same frequency.

Timbral features are measured as Zero Crossing Rate, Spectral Centroid, Roll-off, Brightness, and Cepstral Features.

Zero-crossing rate is defined as the number of sign changes of a signal in a certain period of time. Sign change is defined as the transition of the signal between negative and positive values.

Spectral centroid is a feature used to understand the brightness of sound.

Roll-off is used to approximate the maximum frequency in a signal [24,25].

The feature extraction methods are used to remove irrelevant data from a voice signal. Feature extraction methods are divided into two groups. The criterion for split is the type of feature to be extracted: Spectral feature analysis approaches, which use the speech signal's spectral representation. In addition, temporal feature analysis approaches use a signal's original form [20]. The fundamental operation of feature extraction includes spectrum analysis, parametric transformation, and statistical modeling.

This paper explored some of the feature extraction methods with respect to sounds from musical instruments.

MFCC (Mel-frequency Cepstral Coefficients)

MFCCs are frequently utilized for voice recognition, audio similarity assessment, speech augmentation, music information retrieval, speaker recognition, music genre classification and vowel detection. MFCCs are produced from the cepstral representation of audio clips. MFCCs use the discrete cosine transform of the log power spectrum on a non-linear mel scale to describe the short-time power spectrum of an audio clip. MFCCs are widely used in audio signal processing due to their evenly spaced frequency bands on the mel-scale, closely resembling the auditory system in humans [19]. The MFCC is determined by first dividing the speech signal into alternating frames of 25 or 30 milliseconds each, with a 10-millisecond gap between successive frames. MFCC is well-known and frequently used in the field of voice recognition, however, it has several drawbacks. The fundamental disadvantage of MFCC is its limited resilience against noise signals [19].

Linear Predictive Coding (LPC)

In Linear Predictive Coding (LPC) analysis, a sound sample approximates a linear mixture of prior speech samples. LPC is a frame-based analysis of speech signals. Adjacent frames from the input voice stream are separated and framed as samples. To reduce signal discontinuities, each frame is windowed. This is followed by auto correlating each frame of the windowed signal and transforming each frame of autocorrelations into an LPC parameter set using the

Durbins method [20]. The LPC feature vector was then generated.

Wavelet Packet Decomposition (WPD)

WPD is essentially DWT generalized. WPD is hence more adjustable. As with DWT, WPD is divided into low and high frequency components. In WPD, the transform step is applied to both low and high-pass results, but in DWT, it is only applied to the low pass results. WPD repeatedly decomposes both the approximation and detail coefficients at each level, resulting in a tree-like structure. Each node in the decomposition tree represents a specific frequency band of the signal. The leaf nodes correspond to the finest-scale frequency bands [23].

Perceptual Linear Prediction (PLP)

LPC is a powerful technique for modeling and representing speech signals. Its applications in speech coding, recognition, and identification have made it an essential tool in the field of audio signal processing. It aims to extract features that are more perceptually relevant than those extracted by traditional LPC. At this point, the Hamming Window is utilized, and the power range of the windowed sign is set to FFT [25].

Discrete Wavelet Transform (DWT)

Speech data can be effectively extracted using DWT, which is highly appropriate for non-stationary signal data. For voice feature extraction, it performs better and is more computationally effective and efficient. It is effective across all frequency bands because of its variable window sizes. Two filters, a high-pass filter and a low-pass filter, are applied to the signal, resulting in the production of two signals [25].

5. General Structure of Proposed Model

In this paper, the authors have intended to propose a deep learning-based model to analyse the sounds of the diverse musical instruments as depicted in Figure 2.

The flow of the proposed model is described stepwise as follows

Step 1: A background study of the Indian Instrument Family is done thoroughly in the early stages. In this step, various instruments are identified and analyzed from various categories.

Step 2: Different parameters required for the analysis of musical instruments are recognized in this step.

Step 3: After that available datasets for musical instruments are identified and studied.

Step 4: After the adaption of the dataset the next step is to extract the necessary characteristics

in terms of required features which is named as the process of feature extraction.

Step 5: Once the extraction of the features is done the data set is partitioned as per the requirement of the model creation.

Step 6: Finally, predictions are outlined as percentages and graphs.

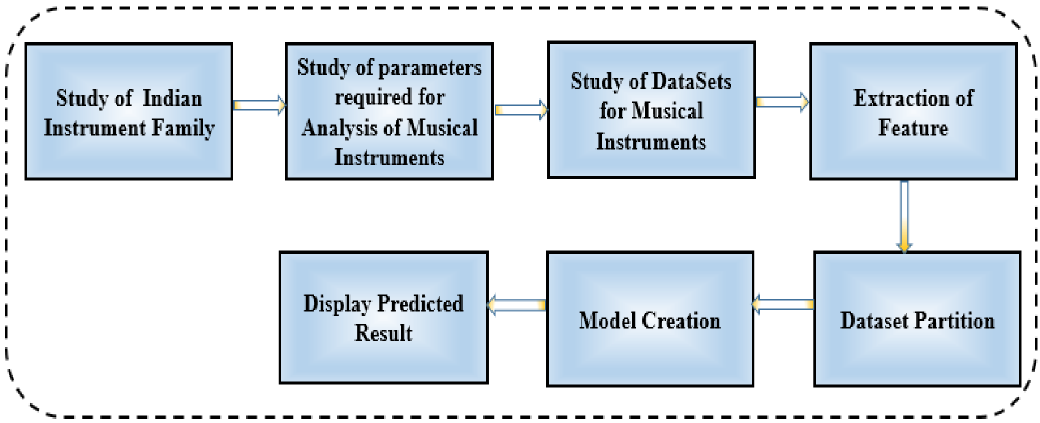
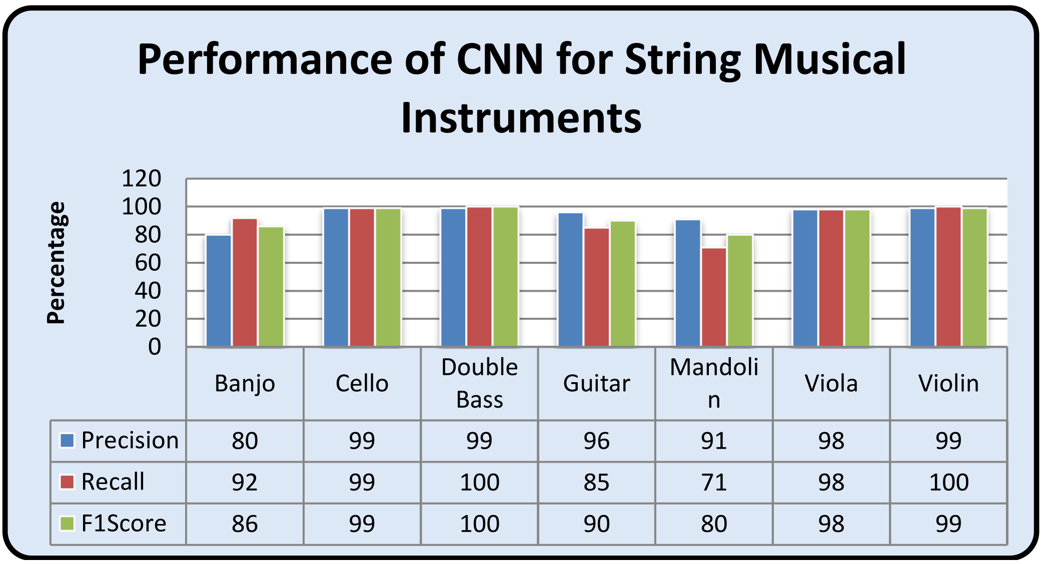


Figure 2. General Structure of Proposed Model for classification of musical instrument sound.

6. Results and Discussions

In this study, by applying the above steps mentioned in section V deep learning architecture CNN is implemented using different Python libraries. The musical string instruments data set required for model building and testing is extracted from Kaggle. Multiclass classification along with 7 classes of string musical instruments is implemented. Results are evaluated in the form of evaluation metrics viz precision, recall, and F1-score to assess the model performance against instruments like banjo, cello, double bass, guitar, mandolin, viola, and violin. Performance metrics commenced by the CNN model to develop a multi-class classifier is sketched in below Figure 3.



	Banjo	Cello	Double Bass	Guitar	Mandolin	Viola	Violin
Precision	80	99	99	96	91	98	99
Recall	92	99	100	85	71	98	100
F1Score	86	99	100	90	80	98	99

Figure 3. Performance of CNN for 7 classes of different String based Musical Instruments

The overall performance of the CNN model is depicted in the below figure 4 in terms of Accuracy, Macro Avg F1 Score and Weighted Avg F1 Score which is 98%, 93% and 98% respectively.

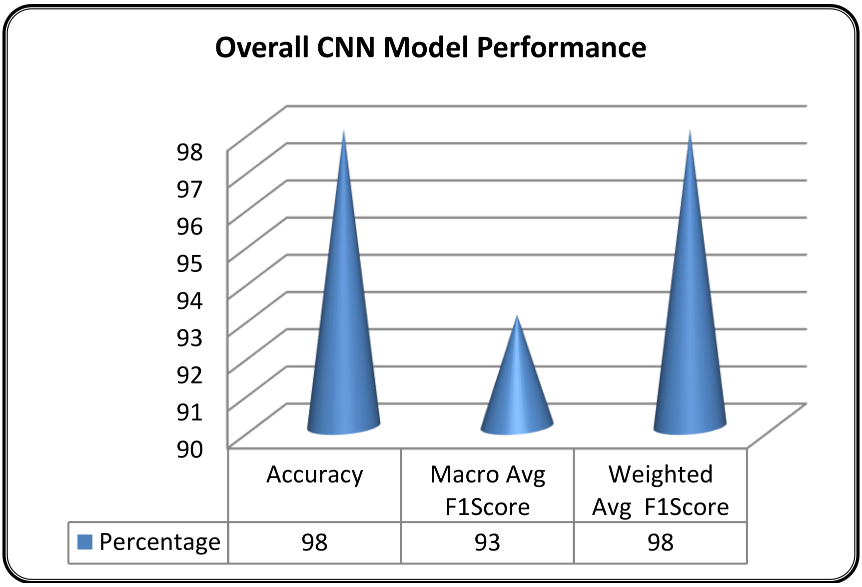


Figure 4. Performance Measures by CNN

7. Conclusion

In the previous several decades, audio signal processing has advanced tremendously in terms of signal analysis and categorization. Music, speech, and background sounds are all included in the audio signal. In this paper, this article focused on the process of analysis of the sounds of instruments in the form of audio signals. The comprehensive structure of the suggested sound classification model for musical instruments is described here for signal analysis, property extraction, behavior prediction, and pattern recognition. We have focused on various feature extraction techniques in this paper. The authors want to explore numerous musical instruments and attributes using machine and deep learning technologies. As a result, the accuracy of the findings may vary depending on the algorithm. In this study, the deep learning algorithm CNN is implemented per the proposed system architecture and predictions are outlined. The CNN model's overall accuracy across all classes is 98%, reflecting its ability to make correct predictions. Also, the performance of CNN against 7-string musical instruments is given in the form of precision, recall, and F1-score which varies as per the instrument. Predictions made by implementing the proposed model might be useful to professionals in the music industry.

References

[1] Constantinescu, C., & Brad, R. (2023). An Overview on Sound Features in Time and Frequency Domain.

International Journal of Advanced Statistics and IT&C for Economics and Life Sciences, 13(1), 45–58. https://doi.org/10.2478/ijasitels-2023-0006
[2] Gulhane Sushen R, Badhe Sanjay S, Dr. Shirbahadurkar Suresh D.(2018), Cepstral (MFCC) Feature and Spectral (Timbral) Features Analysis for Musical Instrument Sounds. Global Conference on Wireless Computing and Networking (GCWCN), Pp. 109-113.
[3] Anuz, H., Masum, A. K. M., Abujar, S., & Hossain, S. A. (2021). Musical Instrument Classification Based on Machine Learning Algorithm. Lecture Notes in Networks and Systems, 164, 57–67. https://doi.org/10.1007/978-981-15-9774-9_6
[4] Prabavathy*, S., Rathikarani, V., & Dhanalakshmi, P. (2020). Classification of Musical Instruments using SVM and KNN. International Journal of Innovative Technology and Exploring Engineering, 9(7), 1186–1190. https://doi.org/10.35940/ijitee.G5836.059720
[5] Xiu CHIAH, C., Choo TAY, L., & Kin LAI, W. (2023). Sensors & Transducers Classifying Musical Instrument Using Spatiotemporal Features with Deep Neural Networks. In Sensors & Transducers (Vol. 263). http://www.sensorsportal.com
[6] Agostini, G., Longari, M., & Pollastri, E. (n.d.). CLASSIFICATION WITH SPECTRAL FEATURES.
[7] Blaszke, M., & Kostek, B. (2022). Musical Instrument Identification Using Deep Learning Approach. Sensors, 22(8). https://doi.org/10.3390/s22083033
[8] Solanki, A., & Pandey, S. (2022). Music instrument recognition using deep convolutional neural networks. International Journal of Information Technology (Singapore), 14(3), 1659–1668. https://doi.org/10.1007/s41870-019-00285-y
[9] Joshi D, Pareek J, Ambatkar P (2023) Comparative Study of Mfcc and Mel Spectrogram for Raga Classification Using CNN. Indian Journal of Science and Technology 16(11), Pp. 816-822. https:// doi.org/10.17485/IJST/v16i11.1809.
[10] Monica S. Nagawade, Varsha R. Ratnaparkhe(2017), Musical Instrument Identification using MFCC. IEEE International Conference On Recent Trends in Electronics Information & Communication Technology (RTEICT), Pp. 2198-2202.
[11] Sarfaraz Masood, Sarfaraz Masood(2015), Novel Approach for Musical Instrument Identification Using Neural Network, IEEE INDICON, Pp. 1-5.
[12] Seyed Muhammad Hossein Mousavi, V. B. Surya Prasath, Seyed Muhammad Hassan Mousavi (2019), Persian Classical Music Instrument Recognition (PCMIR) Using a Novel Persian Music Database, 9th International Conference on Computer and Knowledge Engineering, Pp. 122-133.
[13] D. M. Chandwadkar,M. S. Sutaone(2013), Selecting Proper Features and Classifiers for Accurate Identification of Musical Instruments, International Journal of Machine Learning and Computing, Vol. 3, No. 2, DOI: 10.7763/IJMLC.2013.V3.296, Pp.172-175.
[14] Kumar, V., Recupero, D. R., Riboni, D., & Helaoui, R. (2021). Ensembling Classical Machine Learning and Deep Learning Approaches for Morbidity Identification from Clinical Notes. IEEE Access, 9, 7107–7126. https://doi.org/10.1109/ACCESS.2020.3043221
[15] Norhalina Senan, Rosziati Ibrahim, Nazri Mohd Nawi, Musa Mohd Mokji (2009), Feature Extraction for Traditional Malay Musical Instruments Classification System, International Conference of Soft Computing and Pattern Recognition, Pp.454-459, DOI 10.1109/SoCPaR.2009.94
[16] Garima Sharma , Kartikeyan Umapathy, Sridhar Krishnan(2019). Trends in audio signal feature extraction methods. Published by Elsevier Ltd https://doi.org/10.1016/j.apacoust.2019.107020. Pp-1-21.
[17] Maria Labied, Abdessamad Belangour(2021). Automatic Speech Recognition Features Extraction Techniques: A Multi-criteria Comparison. International Journal of Advanced Computer Science and Applications, Vol. 12, No. 8. Pp.177-182.
[18] Hariharan Muthusamy, Kemal Polat, Sazali Yaacob, 2015, Improved Emotion Recognition Using Gaussian Mixture Model and Extreme Learning Machine in Speech and Glottal Signals, Mathematical Problems in Engineering, Volume 20, Article ID 394083 https://doi.org/10.1155/2015/394083
[19] E.Chandra, K. Manikandan, M. Sivasankar, 2014, A Proportional Study on Feature Extraction Method in Automatic Speech Recognition System, International Journal Of Innovative Research In Electrical, Electronics, Instrumentation And Control Engineering, Vol. 2, Issue 1.
[20] Xuechuan Wang, Kuldip K. Paliwal (2002) A Modified Minimum Classification Error (MCE) Training Algorithm for Dimensionality Reduction, Journal of VLSI Signal Processing 32.
[21] Marko V. Jankovic, Masashi Sugiyama, Probabilistic Principal Component Analysis Based on JoyStick Probability Selector, Available from: https://www.researchgate.net/publication/221533826

[22] Mohammed Arif Mazumder, Rosalina Abdul Salam(2019), Feature Extraction Techniques for Speech Processing: A Review, International Journal of Advanced Trends in Computer Science and Engineering, Vol 8, 1.3, https://doi.org/10.30534/ijatcse/2019/5481.32019.
[23] Qin Lin(2022), Music Score Recognition Method Based on Deep Learning, Hindawi, Computational Intelligence and Neuroscience Volume 2022, Article ID 3022767, 12 pages https://doi.org/10.1155/2022/302276, Pp.1-12.
[24] Constantinescu, C., & Brad, R. (2023). An Overview on Sound Features in Time and Frequency Domain. International Journal of Advanced Statistics and IT&C for Economics and Life Sciences, 13(1), 45–58. https://doi.org/10.2478/ijasitels-2023-0006
[25] Banerjee, A., Ghosh, A., Palit, S., & Ballester, M. A. F. (2018). A novel approach to string instrument recognition. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10884 LNCS, 165–175. https://doi.org/10.1007/978-3-319-94211-7_19
[26] Book [1] Speech and audio processing Dr. Shila D.Apte Wiley India Pvt. Ltd. ISBN:978-81-256-4038-9