Designing Musical Instrument Identification Systems Using Lightweight Machine Learning Architectures

Jashwanth Krishtamaraj[1], Devaraj Jagadiswary[1], Subramani Raja[2], Shubham Sharma[3] and Alaa Kareem Mohammed[4]

[1]*Department of Artificial Intelligence and Data Science, Chennai Institute of Technology, 600069 Chennai, Tamil Nadu, India*

[2]*Center for Advanced Multidisciplinary Research and Innovation, Chennai Institute of Technology, 600069 Chennai, Tamil Nadu, India*

[3]*Department of Mechanical Engineering, Lloyd Institute of Engineering & Technology, Knowledge Park II, Greater Noida, 201306 Uttar Pradesh, India*

[4]*Prosthodontics Techniques Department, Dijlah University College, 10021 Baghdad, Iraq*

sraja@citchennai.net, jagadiswaryd@citchennai.net, shubham543sharma@gmail.com, shubhamsharmacsirclri@gmail.com, alaa.kareem.mohammed@duc.edu.iq

Keywords: Support Vector Machine, Time-Frequency Domain, Voting Classifier, Mel-Frequency Cepstral Coefficients, Persian Instruments.

Abstract: Musicology is a vast field in which the use of electronic instruments has increased significantly with the development of digital audio workstations and virtual instruments. Selecting the appropriate instrument is crucial for producing high-quality music. The traditional approach to identifying musical instruments relies on manually listening to audio recordings, which is time-consuming and inefficient. Therefore, musical signal processing has become one of the most important tasks in audio signal processing. It enables the identification of instruments in music based on their acoustic features. The proposed work focuses on extracting audio features from WAV files and developing an ensemble model that combines Random Forest and Support Vector Machine classifiers using a voting strategy to identify the instruments used in music. The findings highlight the potential applications of this research in music information retrieval systems and contribute to the growing field of computational musicology by providing a scalable and accurate approach to musical instrument identification, achieving an accuracy of 95%.

1 INTRODUCTION

In music, the ability to accurately identify and classify instruments is a critical task, particularly in music retrieval, interactive music games, and audio analysis. With the development of artificial intelligence and machine learning, this process has become more automated and capable of producing more accurate results. Musicians often use reference music to describe the sound they seek; similarly, this work adopts that concept to recognize musical instruments.

In this approach, a combined model consisting of Random Forest and Support Vector Machine analyzes music signals and uses them as a query to recognize instruments whose extracted features are most similar to those of the input audio. Instead of relying on a static library, a train–test strategy is employed, where the model is trained on labeled audio features and tested on mixed audio samples containing different instruments to ensure effective evaluation.

This method facilitates the organization of large music libraries by categorizing audio tracks based on instrument types, enables sound engineers to isolate specific instruments within audio recordings, enhances interactive music games and applications where users can identify or play along with particular instruments, and provides tools for visually impaired individuals to identify instruments in music through auditory cues. To examine the effectiveness of the model, evaluation metrics such as accuracy, precision, and F1-score are utilized. The combined model is also tested on unseen data to assess its adaptability across diverse audio samples.

To determine the most effective feature domain for classification, both time-domain and time–frequency-domain features are extracted and

evaluated. Since time–frequency features demonstrate higher classification accuracy, features such as Mel-frequency cepstral coefficients (MFCCs), chroma, Mel spectrogram, spectral contrast, spectral centroid, tonnetz, and spectral bandwidth are utilized for effective instrument identification.

2 LITERATURE REVIEW

Recent developments in music classification have enabled researchers to explore various techniques to improve classification and identification accuracy. In [1], a dataset of Chinese musical instruments was used to extract MFCC features combined with an attention mechanism using a deep belief network and Bi-GRU, achieving a highest accuracy of 91%. In [5], an automatic music genre classification approach based on texture selection was proposed, demonstrating the use of k-means clustering to capture diverse sound textures within tracks. This method combined texture-level classification through k-means centroids with track-level predictions using majority voting. Four feature sets, including handcrafted and data-driven approaches, were evaluated on benchmark datasets such as GTZAN and Extended Ballroom. The results showed that the k-means texture selection approach outperformed linear down-sampling, achieving improved classification accuracy across multiple datasets.

At the same time, a zero-shot learning framework achieved a precision of 78% using semantic embeddings [9]. An enhanced CNN architecture with duplicated pooling strategies and convolutional layers resulted in an accuracy of 90% [10], whereas in [6], an ensemble voting model combined with MFCC and Mel spectrogram features achieved an accuracy of 56.4%. In [7], CNN, logistic regression, and ANN models were applied using spectrograms and CSV-based audio features; CNN achieved the highest accuracy of 88.5%, followed by ANN with 64.5% and logistic regression with 60.89%.

In [8], a class-conditional embedding model for music source separation was proposed using Gaussian Mixture Models with various covariance structures and compared against a BLSTM-based mask inference baseline. Using the MUSDB18 dataset, the model demonstrated notable improvements in source-to-distortion ratio, with the tied spherical covariance GMM yielding the best performance. In [11], SVM and KNN classifiers were applied using MFCC and sonogram features from a dataset of 1,284 samples across 16 musical instruments. The SVM model achieved the highest accuracy of 99%, while KNN using MFCC features attained 98.22%. In [13], SVM and KNN models achieved a success rate of 91.27% for the pizzicato family using QDA and 93.07% for the sustained family.

Despite these advancements, several gaps remain in musical instrument classification. One major issue is the imbalance or limited size of datasets, which can be addressed using data augmentation techniques such as pitch shifting and SMOTE [16]. Another limitation is the separate use of time-domain and frequency-domain features, which can reduce model robustness. This can be mitigated by incorporating time–frequency features such as chroma and spectral centroid. Furthermore, deep learning models often suffer from a lack of interpretability, making it difficult to understand feature contributions. Utilizing simpler machine learning models or developing interpretable neural architectures can enhance transparency and trust in real-time applications [17]. Real-time deployment also remains challenging due to limited computational resources in embedded and mobile environments; however, this can be improved through optimization techniques such as model pruning and knowledge distillation..

3 AUDIO FEATURES

Audio features play a crucial role in musical information analysis. By analyzing the characteristics of instruments through their audio features, it becomes possible to differentiate one instrument from another. To provide a clearer understanding, Table 1 outlines the extracted feature domains and their descriptions, while Figures 1a, 1b, and 1c visualize audio signals in the time, frequency, and time–frequency domains, respectively.

The following audio features are extracted from the signal and used to identify the instrument being played.

3.1 Chroma

Focuses on the harmonic content of the signal, specifically the 12 unique pitch classes. It represents the energy distribution across the pitches, emphasizing the harmonic structure. Chroma divides the frequency spectrum into frames and maps each frequency bin to one of the 12 pitch classes. The harmonics of the fundamental frequency are summed within each scale. Since Chroma sums energy over the pitch scale, it is invariant to octave, making it effective for capturing harmonic content [1].

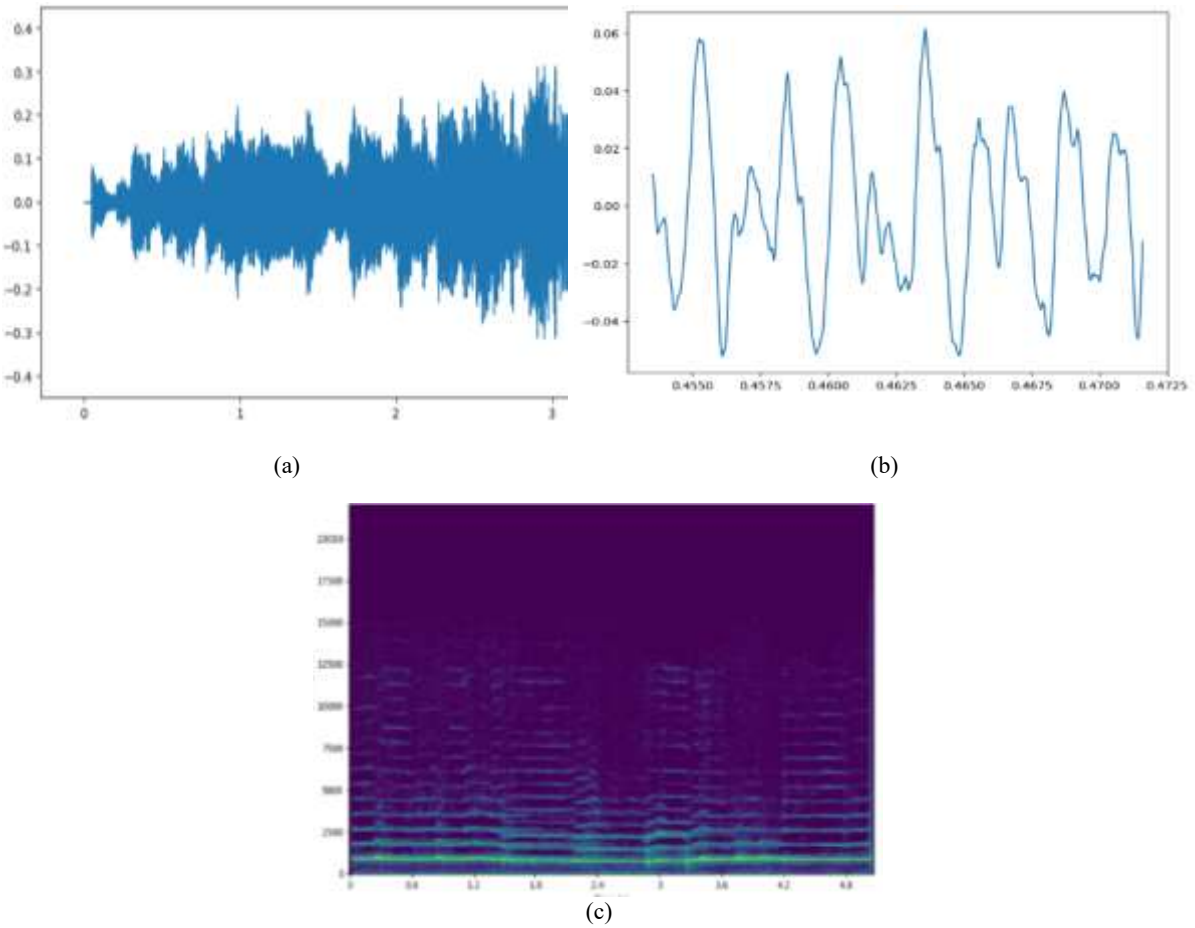


Figure 1: Visualization of audio signals in different domains: a) Time domain, b) Frequency domain, c) Time-Frequency domain.

3.2 Spectral Centroid

Measures the "center of mass" of the spectrum, indicating the brightness of a sound. A higher centroid corresponds to higher perceived brightness. Instruments like the violin have higher spectral centroid values, while instruments like the double bass have lower values, helping to distinguish between them.

3.3 Spectral Rolloff

Captures how the spectral energy is distributed across frequencies. Drums and guitars typically have high rolloff values, whereas instruments like the flute have lower values. When combined with other features, spectral rolloff contributes to effective instrument classification [12].

3.4 Mel-Frequency Cepstral Coefficients (MFCCs)

Capture the timbre of the signal. The signal is divided into short frames, transformed into the frequency domain using the Fast Fourier Transform (FFT), and mapped onto the Mel scale, which mimics human auditory perception. The discrete cosine transform is then applied to reduce dimensionality, resulting in a compact, manageable representation of the audio [2].

3.5 Mel Spectrogram

A time-frequency representation that captures both temporal and spectral information, mapping the frequency axis to the Mel scale. The signal is windowed into small frames, FFT is applied, and frequency bins are mapped to the Mel scale to emphasize lower frequencies. The resulting matrix shows energy distribution over time, helping to focus

on frequencies important for distinguishing instrument sounds.

3.6 Spectral Contrast

Measures the difference between peaks and valleys in each frequency band, indicating harmonic richness. The spectrum is divided into multiple bands, and the difference between peak and minimum energy in each band is calculated. This helps differentiate harmonically rich instruments (e.g., clarinet) from noisier instruments [15].

3.7 Tonnetz

Represents the tonal evolution of the signal, highlighting the overall harmonic structure. Tonnetz features are computed by analyzing the relationships between frequency peaks and their harmonics, helping to distinguish instruments based on changes in harmonic content over time [14].

3.8 Spectral Bandwidth

Measures the spread of the spectrum around the spectral centroid, providing insight into the brightness or dullness of the sound. It indicates how much the frequency content deviates from the centroid, highlighting harmonic richness and helping to distinguish between instruments with broad harmonic spectra versus those with narrow, pure tones.

4 PROPOSED METHODOLOGY

This research follows a systematic way as illustrated in Figure 1. Initially musical audio of 13 instruments is collected from [4], where each instrument has 30 audio files. Audio of instruments such as Flute, Clarinet, Violin, Cello, Saxophone, Double bass, Bass draw and Persian instruments such as Ney, Tar, Santur, Setar, Kamancheh, Hihat data is collected is collected from [3] for diverse data purposes, and then pre-processed by converting mp3 files into wave files for effective feature extraction and to avoid bias and increase effectiveness files are made to have 5 seconds audio. Classification models such as Random-Forest classifier and Support Vector Machine have been combined as an ensemble model, and their results are evaluated using evaluation metrics were discussed. The various steps involved in the proposed system are as Figure 2.

Figure 2: Methodology.

Preprocessing: In the preprocessing stage, the data set contains raw musical audio of 13 distinct instruments. To enhance the identification process, essential pre-processing steps like trimming and padding are implemented. Audio files exceeding 5 seconds are trimmed to have precisely 5 seconds and files shorter than 5 seconds are padded at the end with 0 to ensure uniform length across all audio clips and to avoid unwanted error and bias problems. The mathematical expression of preprocessing is given in equation 1 as follows:

$$\text{Pre-processed(x)} = \begin{cases} Trim(x,5) \;\; if\; L(x) > 5 \\ Pad(x,5)\; if\; L(x) < 5 \end{cases}. \quad (1)$$

Feature Extraction: Using python script along with librosa library acoustic features from the audio file is extracted as numerical representation for computational analysis. Time domain and time-frequency domain features are extracted independently to ensure that different characteristics of the instrument is collected. These extracted features are then organised and stored in a dataframe, making them readily available for further analysis. Figure 3 visualises the process of feature extraction.

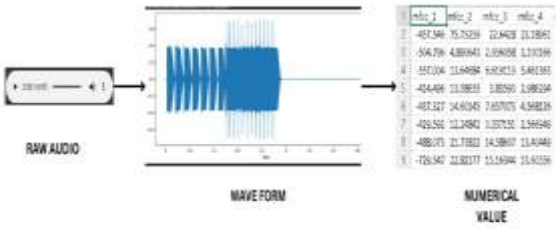


Figure 3: Process of feature extraction.

5 PROPOSED MODEL

The designed model is built using a hybrid approach by combining the architecture of random forest and support vector machine as shown in Figure 4. The process begins with feature extraction of 13 instruments on time and time frequency domain. Features are stored in CSV file and then the data set is divided into training and testing sets. A grid search is utilized to optimize model's parameters, For the

SVM, the grid search utilizes parameters such as penalty parameter C, kernel types such as linear and radial basis function and gamma values. Similarly, for the RF, grid search tunes number of estimators, maximum tree depth, and minimum samples for node splitting.

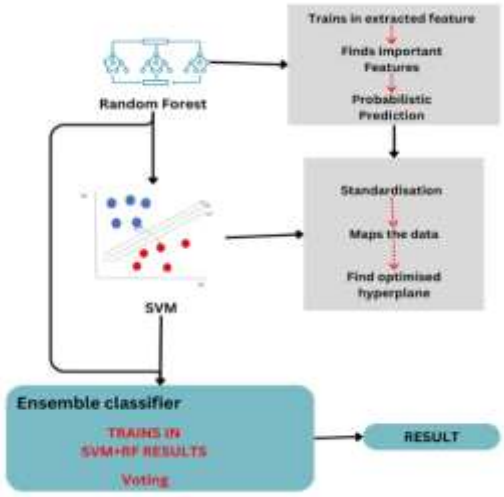


Figure 4: Model work flow.

Once the optimal hyperparameters are computed rf classifier is trained on the extracted data. Then probabilistic outcome denotes the confidence of rf in the classification which is then combined with original features and used as input for SVM model. Since SVM is sensitive to feature scaling, the combined feature set is standardized using a Standard Scaler to ensure all features contribute equally to the model's performance. The SVM is then trained on the standardized combined features, leveraging the RF probabilities as additional inputs to clarify classification boundaries.

At the final step, weights are assigned to both model's result and a voting classifier is built to enhance the classification process. With the strength of rf's probabilistic prediction and with decision making strength of SVM, ensemble model shows a robust performance. During evaluation, the ensemble achieves an accuracy of 95%, outperforming individual classifiers and showcasing its efficiency in multi-class musical instrument classification. This systematic integration of RF and SVM highlights the benefits of combining diverse machine learning algorithms for complex identification tasks.

6 RESULTS AND DISSCUSION

This section showcases the results of combination of RF + SVM and ensemble model in both time domain features and time-frequency domain features and then results such as model's accuracy, precision, recall and F1 score are discussed along with confusion matrix. This enables to analyse the model's strength and weakness.

Time Domain. In time domain analysis, this work evaluated the performance of two models incorporating RF, SVM and voting classifier. As shown in Table 1, the SVM + RF model achieved an accuracy of 0.41, a precision of 0.43, and an F1 score of 0.39, indicating moderate classification efficiency. However, ensemble model built to increase the efficiency under performs with an accuracy of 0.37, precision of 0.41, and F1 score of 0.37.

Table 1: Cases the efficiency of models in time domain features.

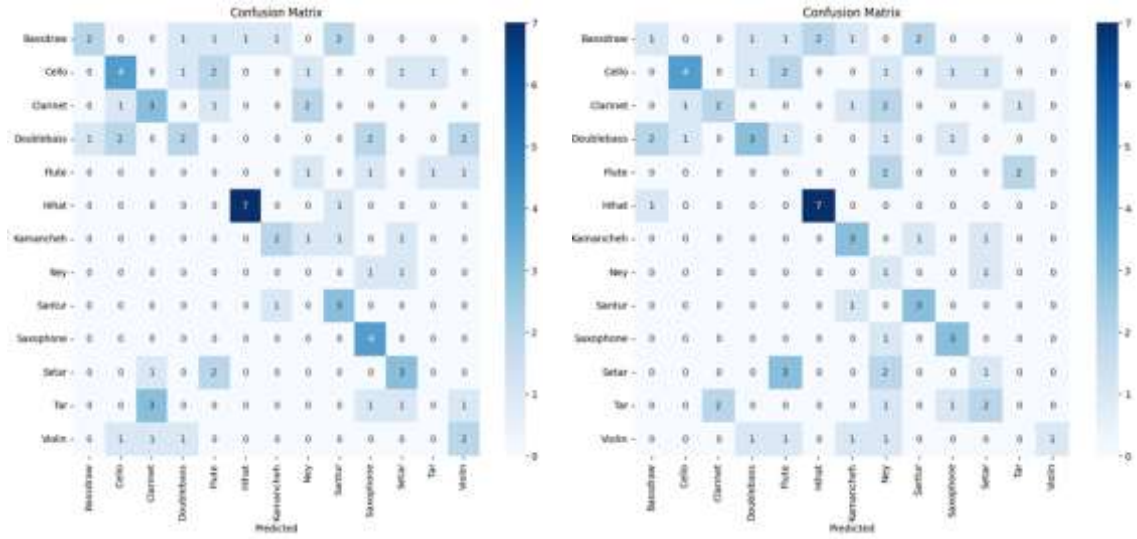
Method	Accuracy	Precision	F1 Score
SVM+ RF	0.41	0.43	0.39
SVM + RF+ Voting classifier	0.37	0.41	0.37

The confusion matrix in Figure 5a and Figure 5b highlights sparse diagonal values and more off-diagonal values, which indicates frequent misclassifications. These results suggest that only time domain features are not much effective for diverse musical instrumental classification.

Table 2 compares the efficiency of models in the time-frequency domain. The SVM + RF method achieves an accuracy of 0.90, precision of 0.91, and F1 score of 0.89. Enhancing this with a Voting Classifier improves performance by achieving an accuracy of 0.95, precision of 0.95, and F1 score of 0.94. This indicates that the Voting Classifier effectively enhances model performance.

Table 2: Highlights the efficiency of models in time-frequency domain.

Method	Accuracy	Precision	F1 Score
SVM+ RF	0.90	0.91	0.89
SVM + RF +Voting classifier	0.95	0.95	0.94



(a) (b)

Figure 5: Comparison of Confusion matrices for different models: a) Confusion matrix for SVM+RF, b) Confusion matrix for ensemble model.

The precision, recall and F1 score emphasize the efficiency of the models, especially hihat, kamancheh and santur with perfect score of 1.00 for all metrics as shown in Table 3. At the same time violin has recall of 0.60 and F1 score of 0.75, which shows model fails to classify it accurately. For instruments like tar, clarinet, double bass model performs with high efficiency. Overall, the SVM + RF model performs better with some future enhancements.

Table 3: Highlights the precision, recall and F1 score of SVM+RF approach.

Instruments	Precision	Recall	F1 Score
Bass draw	0.80	1.00	0.89
Cello	0.90	0.90	0.90
Clarinet	1.00	0.71	0.83
Double bass	1.00	0.78	0.88
Flute	0.67	1.00	0.80
Hihat	1.00	1.00	1.00
Kamencheh	1.00	1.00	1.00
Ney	0.67	1.00	0.80
Santur	1.00	1.00	1.00
Saxophone	0.80	1.00	0.89
Setar	1.00	0.83	0.91
Tar	0.86	1.00	0.92
Violin	1.00	0.60	0.75

The performance of support vector machine + random forest is showcased used confusion matrix and evaluation metrics based on train, test validation. High values among the diagonal in confusion matrix Figure 6 shows that model accurately classifies most of the instruments, while few non diagonal values highlights that the model failed to accurately classify certain instruments like ney, flute and setar.

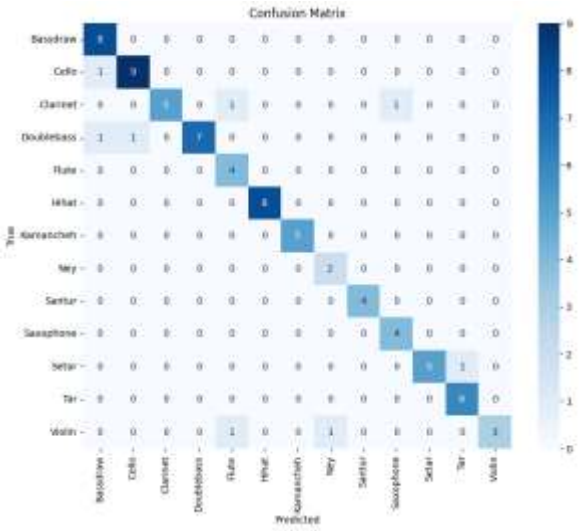


Figure 6: Confusion matrix for RF+SVM.

The high evaluation metrics for the ensemble model (RF+ SVM+ Voting Classifier) highlights the effectiveness of model in classifying instruments, especially for instruments like flute, ney, Hihat, Kamancheh and clarinet the model achieved perfect score of 1.00 as shown in Table 4 for all, which indicates that model perfectly classifying these

models without any false positives and false negatives. The model also well performed for instruments such as double bass, cello and violin with slightly lower F1 score range between 0.83 to 0.95 and also slightly less for recall. These indicates that the identification of these instruments might have under- identification. Confusion shows the performance of model by displaying the correct and incorrect predictions across all the classes.

Table 4: Spotlight the evaluation metrics for RF + SVM + Voting classifier.

Instruments	Precision	Recall	F1 Score
Bass draw	1.00	1.00	1.00
Cello	0.91	1.00	0.95
Clarinet	1.00	1.00	1.00
Double bass	1.00	0.89	0.94
Flute	1.00	1.00	1.00
Hihat	1.00	1.00	1.00
Kamencheh	1.00	1.00	1.00
Ney	1.00	1.00	1.00
Santur	0.80	1.00	0.89
Saxophone	0.80	1.00	0.89
Setar	1.00	0.83	0.91
Tar	0.83	0.83	0.83
Violin	1.00	0.80	0.89

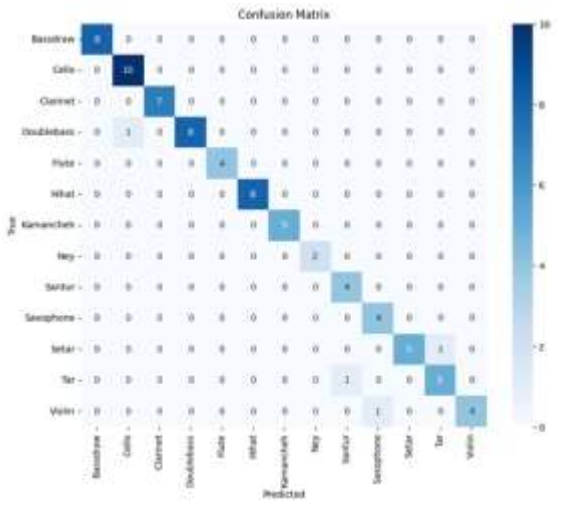


Figure 7: Confussion matrix of ensemble model.

In Figure 7 diagonal values are high indicating that most predictions align with true classes. Additionally, four off- diagonal entries have value of 1, shows that model classified majority of classes accurately.

Figure 8 highlights the accuracy of models in identifying musical instruments using two domains time and time frequency domain individually. It illustrates that, model leveraging time frequency domain features perform better than those only uses time domain, and also in specific ensemble model in TF domain performs better than all models.

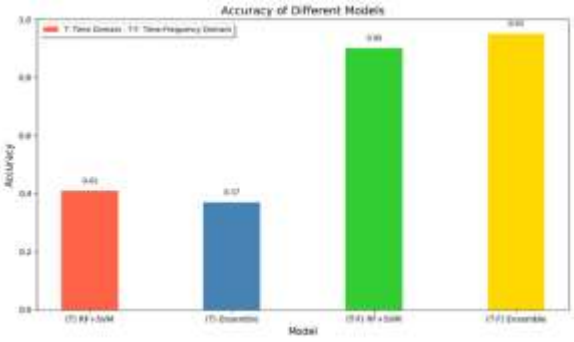


Figure 8: Accuracy of multiple methods.

7 CONCLUSIONS

This paper presents an ensemble method for musical instrument recognition using Random Forest, Support Vector Machine, and a voting classifier. By extracting various audio features, the proposed method achieved an accuracy of 95% on time-frequency domain features, demonstrating its potential for capturing instrumental information from music. The model has shown robustness and adaptability across different datasets, indicating its applicability in music production, interactive applications, accessibility tools, and identification of both traditional and modern instruments based on their acoustic characteristics. Future work could involve incorporating new machine learning techniques, expanding the range of audio features, and including more diverse datasets to further enhance efficiency, accuracy, and generalization. The method's ability to capture nuanced audio signatures makes it a valuable tool for detailed musical analysis and cataloging. The results also highlight its potential for integration into real-time systems for live performance or automated music tagging.

8 FUTURE WORKS

Musical instrument identification faces several challenges, such as limited-quality datasets, restricted diversity of instruments, and models struggling to classify unseen data due to noise and insufficient training samples, which can lead to generalization issues. Additionally, most approaches rely solely on

audio features and do not incorporate visual information from images or videos of instruments, which could further enhance model performance. Another difficulty arises when recognizing instruments in polyphonic or heavily layered compositions, where overlapping sounds can obscure the characteristics of individual instruments.

To address these limitations, future research could focus on methods for effective spatial feature extraction and the use of multimodal data, combining acoustic and visual information about instruments. Leveraging unsupervised and semi-supervised learning techniques, such as clustering and autoencoders, can enable classification without complete dependence on labeled data. Efforts to improve robustness against noise and acoustic variability are essential, along with the development of techniques for identifying instruments in polyphonic and heavily layered compositions. Incorporating diverse datasets, variations in noise, different playing techniques, and a wider range of instrument usage can further enhance the performance and generalizability of musical instrument identification methods.

ACKNOWLEDGEMENT

This work is partially funded by Center for Advanced Multidisciplinary Research and Innovation. Chennai Institute of technology, India, vide funding number CIT/CAMRI/2025/CFR/010.

REFERRENCES

[1] G. Sharma, K. Umapathy, and S. Krishnan, “Trends in audio signal feature extraction methods,” Applied Acoustics, vol. 158, art. no. 107020, 2020.

[2] M. K. Gourisaria, R. Agrawal, M. Sahni, and P. K. Singh, “Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques,” Discover Internet of Things, vol. 4, no. 1, p. 1, 2024.

[3] S. M. H. Mousavi, V. B. S. Prasath, and S. M. H. Mousavi, “Persian classical music instrument recognition (PCMIR) using a novel Persian music database,” in Proc. 9th Int. Conf. Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 2019, pp. 122–130, doi: 10.1109/ICCKE48569.2019.8965166.

[4] Seth814, “Dataset for audio classification,” GitHub repository, 2018. [Online]. Available: https://github.com/seth814/Audio-Classification/tree/master/wavfiles.

[5] J. H. Foleis and T. F. Tavares, “Texture selection for automatic music genre classification,” Applied Soft Computing, vol. 89, art. no. 106127, 2020.

[6] D. Kostrzewa, P. Kaminski, and R. Brzeski, “Music genre classification: Looking for the perfect network,” in Proc. Int. Conf. Computational Science, Cham, Switzerland: Springer, June 2021, pp. 55–67.

[7] S. Chillara, A. S. Kavitha, S. A. Neginhal, S. Haldia, and K. S. Vidyullatha, “Music genre classification using machine learning algorithms: A comparison,” Int. Research Journal of Engineering and Technology, vol. 6, no. 5, pp. 851–858, 2019.

[8] P. Seetharaman, G. Wichern, S. Venkataramani, and J. Le Roux, “Class-conditional embeddings for music source separation,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 301–305.

[9] J. Choi, J. Lee, J. Park, and J. Nam, “Zero-shot learning for audio-based music classification and tagging,” arXiv preprint, arXiv:1907.02670, 2019.

[10] H. Yang and W. Q. Zhang, “Music genre classification using duplicated convolutional layers in neural networks,” in Proc. Interspeech, Sept. 2019, pp. 3382–3386.

[11] S. Prabavathy, V. Rathikarani, and P. Dhanalakshmi, “Classification of musical instruments using SVM and KNN,” Int. J. Innovative Technology and Exploring Engineering, vol. 9, no. 7, pp. 1186–1190, 2020.

[12] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, pp. 293–302, 2002.

[13] G. Agostini, M. Longari, and E. Pollastri, “Musical instrument timbres classification with spectral features,” EURASIP Journal on Advances in Signal Processing, vol. 2003, pp. 1–10, 2003.

[14] D. H. Rudd, H. Huo, and G. Xu, “Leveraged mel spectrograms using harmonic and percussive components in speech emotion recognition,” in Proc. Pacific-Asia Conf. Knowledge Discovery and Data Mining, Cham, Switzerland: Springer, May 2022, pp. 392–404.

[15] S. Kumar and S. Thiruvenkadam, “An analysis of the impact of spectral contrast feature in speech emotion recognition,” Int. J. Recent Contributions in Engineering, Science & IT, vol. 9, no. 2, pp. 87–95, 2021.

[16] O. O. Abayomi-Alli, R. Damaševičius, A. Qazi, M. Adedoyin-Olowe, and S. Misra, “Data augmentation and deep learning methods in sound classification: A systematic review,” Electronics, vol. 11, no. 22, art. no. 3795, 2022.

[17] P. Zinemanas, M. Rocamora, M. Miron, F. Font, and X. Serra, “An interpretable deep learning model for automatic sound classification,” Electronics, vol. 10, no. 7, art. no. 850, 2021.