Deep Neural Network and SVM-Based Multiclass Musical Instruments Classification Using MFCC, STFT, and Related Acoustic Features

Yu Liu
Music education, Xi 'an Shiyou University, 710065, Shaanxi, China
E-mail: chenghui710@163.com

Keywords: artificial intelligence, deep learning, audio signals, musical instruments, classification, waveforms

Artificial Intelligence (AI) has revolutionized the field of music analysis by enabling advanced sound recognition and classification techniques. In the recent year, the music industry has had transformative evolution in recent years, significantly impacting user engagement, creativity, and technological innovation across various domains, including entertainment, education, and therapy. Musical instrument recognition is an emerging field within this landscape that could be used in applications such as automated music transcription and intelligent recommendation systems and adaptive music generation. Using Deep Learning (DL) alongside Artificial Intelligence has dramatically improved how we hear things as it has become a robust analytics tool for patterns in audio. The intricate signals from audio data together with overlapping frequencies and instrument diversity present significant challenges to accurate musical instrument prediction and classification. Traditional structured machine learning models cannot be applied successfully to dealing with complicated patterns, in contrast DL models possess superior designed system architectures and stronger feature extracting ability. Through the integration of these DNN with default layers with support vector machines, an instrument recognition framework is presented in this research, on publicly available dataset of diverse instruments with three second clips combined with Mel Spectrogram and its audio features. The standard measures are used for measuring performance of models such as accuracy, precision, recall and f1-score. The proposed model DNN achieves 98% classification precision over the SVM baseline with accuracy of 96% using musical instrument dataset with 24 classes. Using this research as the base of concept, it is shown that DL is better than current proposed methodologies at improving audio transformation processes, and it promises potential in improving the state of the art at musical instrument identification techniques that would yield useful results to intelligent music systems and AI audio analysis methodologies.

Povzetek: Raziskava izboljša prepoznavanje glasbil z globokim učenjem, uporabo DNN in SVM za razvrščanje glasbenih instrumentov s pomočjo MFCC, STFT in drugih akustičnih lastnosti, ter doseže 98% točnost.

1 Introduction

Artificial Intelligence (AI) has revolutionized the way music is analyzed, categorized, and experienced. With advancements in deep learning and signal processing, AI-driven systems have enabled automatic recognition of musical instruments, improving applications such as music recommendation systems, content-based audio retrieval, and synthesis [1]. Music has been an integral part of human culture and society for centuries, influencing emotions, behaviors, and interactions in profound ways. With the advent of technology, the music industry has undergone significant transformations, particularly in the way music is produced, distributed, and consumed. Music has become more available than ever now in modern trends making music accessible to users and making them explore and enjoy different genres and instrumental music [2]. This leads to the need of understanding and categorizing instrumental sounds for applications, including music recommendation systems [3], content-based audio retrieval and synthesis [4]. These developments have enhanced user experience, and they provide scope for innovation in several areas including entertainment [5], therapy and education [6]. Music categorization is becoming an increasingly relevant issue of research so much that the characteristics of instrumental sounds are better to be understood [7]. Classifying musical instruments from their audio signals has many implications, such as assisting composers, helping musicologists during their analysis, or improving user engagement through personal recommendations [8]. But this is a very challenging task, because the sound patterns are very complex, and different instruments have very intricate timbres and frequencies that overlap. However, with the advancement of Artificial Intelligence (AI) they help to explore and solve these problems not only efficiently but also effectively [9].

Machine Learning (ML) and especially Deep Learning (DL) have unfolded true artificial intelligence revolution in the background of field like delivery systems, mapping, medical research and design practices in the field of music analysis to extract meaningful features from complex signals and result in precise classification of different sounds [10] [11]. These models can capitalize on more sophisticated methods to detect sophisticated variances in audio traits, for example pitch, rhythm, and timbre, that are important for separate musical instruments. Nevertheless, the inestimable nature of audio data necessitates elaborate techniques and rich datasets to produce dependable forecasts [12].

In this research, we identify the difficulty of predicting musical instrument sounds using ML and DL models. It highlights the audio feature of waveform, Chroma Short Time Fourier Transform (STFT), Zero crossing (ZCR), Mel Frequency Cepstral Coefficients (MFCCs), Root Mean Square (RMS) energy and the Mel Spectrogram for classification of instrumental audio. To realize these, models including Support Vector Machines (SVM) and Deep Neural Networks (DNN) were implemented with a publicly available online dataset. With these techniques applied, the study attempts to present a reliable and powerful method of instrument sound classification.

- *Development of a Predictive Framework:* Designed and implemented a framework in the process of applying machine learning (SVM) and deep learning (DNN) models achieving highest performance of 98% for musical instrument classification.
- *Utilization of Diverse Audio Features:* The audio feature included in this research is the waveform, Chroma Short-Time Fourier Transform (STFT), Zero Crossing Rate (ZCR), Mel Frequency Cepstral Coefficients (MFCCs), Root Mean Square (RMS) energy and Mel Spectrogram.
- *Contribution to Music Technology:* It enhanced our understanding and methodology for sound-based classification, which is the basis for the application in music recommendation systems, content-based audio collection, and synthetic sound.

The remainder of this paper is organized as follows: Section 2 covers related work, discussions of previous studies and methodologies of musical instrument classification. Problem statement discussed in Section 3. Research methodology is described in Section 4 that includes dataset selection, feature extraction techniques, and ML and DL model implementation. The results and discussion are presented in Section 5 as we analyze the performance of the models and their ability to predict instrumental sounds. Section 6 shares the discussion analysis based on cross findings. Section 7 closes the study with summation of key findings, and suggestions of future research directions.

2 Related work

In recent years, efforts on musical instrument classification have become quite popular using machine learning techniques, as summary defined in table I. The study [13] a wide range of different machine learning methods, such as Naive Bayes, Support Vector Machines (SVM), Random Forests, boosting techniques (such as AdaBoost, XGBoost) and deep learning models (such as Convolutional Neural Networks (CNNs) and Artificial Neural Networks (ANNs)). Finally, the effectiveness of these methods was evaluated on NSynth, showing the benefits and limitations of each method. Further another study [14], were interested in classifying acoustic instruments using CNNs. They extracted features such as Mel spectrograms and Mel Frequency Cepstral Coefficients (MFCCs) of their data, from a dataset on Kaggle, containing audio recordings of piano, violin, drums and guitar. The most beneficial use of a comprehensive feature set for accurate classification that they found. In [15] presented a musical instrument classification algorithm by using multi-channel feature fusion and XGBoost. They input audio features extracted and fused into the XGBoost model for training by extracting it. On the Medley-solos-DB dataset and provides a technique for feature selection in this music instrument classification task. An artificial neural network (ANN) model [16] trained to classify 20 different classes of musical instruments. On the London Philharmonic Orchestra dataset, they only used the MFCCs of the audio data and trained to state-of-the art accuracy. In the first end to end adversarial attacks affecting a music instrument classification system, [17] were able to perform attacks on audio waveforms directly rather than spectrograms. We demonstrate that their attacks reduce accuracy close to a random baseline while preventing even imperceptible perturbations, calling into question such systems' validity. The automatic instrumentation of symbolic multitrack music is a feasible method [18] for learning to separate parts. We treat the task of part separation as a sequential multi-class classification problem and utilize machine learning to map raw notes sequences to part label sequences, beating several baselines. In [19], we used spectrograms as the input to the CNNs used for the recognition of musical instruments, as it captured the local patterns contained in the data. It was shown that deep architectures could learn practical audio features without the manual design of features through the research. The [20] study looking at how sound sounds of instruments could be quantified by harmonic frequency content apparent in spectrograms. To 80% accuracy the researchers were able to propose the use of a simple but efficient K-Nearest-Neighbors machine learning algorithm. It was found that a larger dataset, and using convolutional neural networks, could improve classification accuracy. A study [21] looked at using machine learning to identify different instruments by analyzing the harmonic frequency content in spectrograms. A simple yet effective K-Nearest-Neighbors (KNN) algorithm was proposed by researchers which appears to reach nearly 80% accuracy. Since they recommended classification accuracy can be further improved using larger dataset and CNNs. In [22], built a parallel CNN-BiGRU model for polyphonic instrument classification from raw audio waveforms and achieve competitive results on the IRMAS dataset, which has lower latency due to separability without necessary pre-processing of the input signals. The YOLOv7 was used [23] to recognize similar musical instruments with an average accuracy of 86.7%. In his work [24], classified traditional Chinese musical instruments through a deep belief network. These studies highlight the validity of using deep learning models to learn highly expressive audio features for classification tasks.

In spite of exciting developments in the field of classification of musical instruments, state-of-the-art work shows significant limitations regarding generalization power, feature describing ability and class diversity. Most of existing studies ([13], [14], [19]), using CNN or ensemble methods, exploit constrained datasets, like NSynth or Kaggle subsets, allowing only to put a tight focus on a handful of instruments (usually 4–10 classes). Moreover, the majority of the models also utilize MFCC or Mel Spectrograms as the sole extractor of features neglecting the prospects for combined/complementary features such as STFT, waveform shape, etc., which are essential for differentiating between acoustically similar instruments. Although some high accuracies have been reported in such cases, for instance, 97% with XGBoost on Medley-solos-DB ([15]), this kind of studies usually employs smaller or curated datasets, producing the limited application for more varied real-world scenarios. Furthermore, not many studies examine the robustness of models in situations that may include a highly imbalanced class or a low-sample instrument.

These gaps have been eliminated in this study with a suggested hybrid Deep Neural Network (DNN) and SVM-based framework for classification of the 28 classes of instruments using a rich feature set of MFCC, STFT, and spectral descriptors. In contrast to previous models, the presented approach increases the classification scale by a vast number of instruments types and exhibits improved generalization power in the form of more profound networks and improved training procedures. This extensive configuration seeks to fill the gap between the bench marking for the academics and the implementation of music information retrieval systems.

2.1 Research questions

RQ1: How well acoustic feature-based Deep Neural Network (DNN) would be able to classify 28 musical instrument classes including MFCC, STFT and spectral descriptors?

RQ2: Does the incorporation of STFT with MFCC enhance the classification performance over MFCC based models?

RQ3: How does a DNN model perform as compared to a Support Vector Machine (SVM), in measuring accuracy, generalization, and class imbalance on multiclass instrument classification?

RQ4: How does the effect of the change in network depth, activation functions, and optimization strategy impact the capability of classifying acoustically similar instrument.

Table 1: Summary of existing studies

Ref	Model	Dataset	Features	Results (Acc %)
[13]	NB, SVM, RF, CNN, ANN	NSynth Dataset	Mel Spectrogram, MFCCs	85
[14]	CNN	Kaggle (Piano, Violin, Guitar, Drums)	Mel Spectrogram, MFCCs	89
[21]	KNN	Custom dataset	Harmonic Frequency Content (Spectrograms)	~80%
[15]	XGBoost	Medley-solos-DB	Feature fusion (Waveform, Spectrograms, MFCCs)	97
[16]	ANN	London Philharmonic	MFCCs	82
[17]	Adversarial attacks on CNN models	NSynth Dataset	Waveform perturbations	71
[18]	Sequential Multi-Class Classifier	Symbolic Multitrack Music	Note sequences	87
[19]	CNN, ANN, Random Forests, XGBoost	NSynth Dataset	Mel Spectrograms, MFCCs	79
[20]	CNN	Custom dataset	Spectrogram inputs	88
[23]	YOLOv7	Kaggle data	MFCC	86
[24]	Deep Belief Network	Traditional Chinese Instruments	Waveform, MFCC	90

3 Problem statement and formulation

This study involves identifying and classifying musical instruments from audio signals using machine learning and deep learning models. We are given a dataset of 3 second audio clips representing a single unique musical instrument and asked to develop a framework capable of accurately classifying the instrument type within a dataset based on the extracted audio features. One challenge in all this is the complexity of such different audio characteristics of different instruments: how do you handle polyphonic sounds (where more than one instrument is played at once), how do you make predictions on diverse datasets and across different instrument types. Formally, Let $D = \{d_1, d_2, \dots., d_n\}$ represent a collection of n audio samples, where each sample d_i corresponds to an audio clip of length $t = 3$ seconds and is associated with a label $y_i \epsilon I = \{I_1, I_2, \dots., I_j\}$. The goal is to design a function $f: D \rightarrow I$, such that each input sample d_i, the model correctly predicts, the associated instrument label $f(d_i) = y_i$. The model is to be trained using a combination of various audio features such as Waveform, Chroma STFT, ZCR, MFCCs, RMS, and Mel-Spectrogram, which will be derived from the audio clips. The objective is to maximize the accuracy of the prediction function f across the entire dataset, ensuring effective recognition of musical instruments from new, unseen audio samples.

4 Methodology

A structured approach to analyzing interactive media images data preprocessing, feature engineering and model development is proposed in the research methodology, as shown in fig 1. In the first, raw dataset is passed through extensive preprocessing process, wherein, noise is trimmed, and normalization techniques are applied to make sure that data are consistent and prepared. Utilizing feature engineering, we extract relevant useful acoustic features for the purpose of improving model performance in discriminating between categories. The methodology is to develop a proposed model, to be evaluated and compared with existing approaches. The performance of the proposed model will be compared in terms of its accuracy, precision, recall and other relevant metrics using the comparison-based analysis, which would justify why it is a useful model in classifying interactive musical sounds.

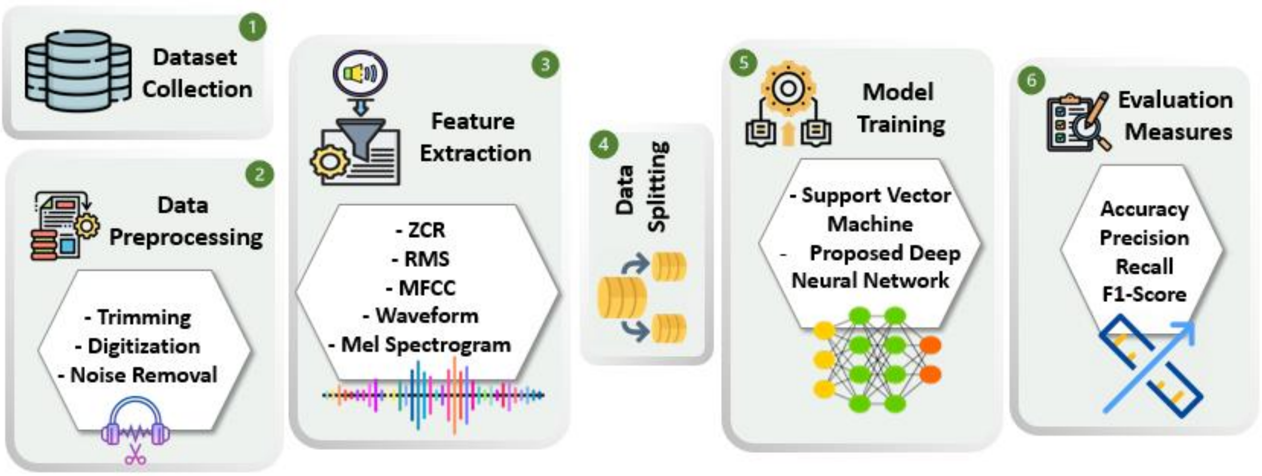


Figure 1: Research proposed methodology

4.1 Data preprocessing

Data preprocessing is one most crucial step when preparing dataset for training machine learning as well as deep learning models. To improve the quality of the audio data and have consistent data across the dataset, several preprocessing techniques were applied in this study. Noise removal was performed first to eliminate any unwanted background noises which might have a negative impact on the model's capability to learn the real characteristics of the musical instruments. The underlying processing rule consists in filtering the frequencies outside the expected range for each instrument and retaining only the right acoustic features. Let $x(t)$ represent the original audio signal and $n(t)$ denote the noise, where the clean signal $s(t)$ is derived by applying a noise removal filter f, such that $s(t) = f\left(x(t)\right) - n(t)$. Next, the audio clips were trimmed to a fixed length of 3 seconds to ensure uniformity. This step involves selecting the first 3 seconds of each audio clip $x(t)$ and discarding any excess duration, ensuring consistency across the dataset as $x_{trimmed}(t)$ represent the trimmed audio as $x_{trimmed}(t) = x(t) \rightarrow t \epsilon$ [0,3]. After that, Padding was then applied to the audio clips, involves appending zeros to the end of clips that are shorter than the required length as $x_{padded}(t) \rightarrow \begin{cases} x(t)\ for\ t\ \epsilon\ [0,T] \\ 0 \quad for\ \ t\ \epsilon\ (T,3] \end{cases}$. Finally, data digitization was applied to convert the continuous audio signals into a digital form suitable for model processing. This involves the continuous audio signal at a fixed rate, resulting in discrete signal. Let $x_{digitized}(n)$ represent the digitized signal, where n is the discrete time index and the signal is obtained by sampling the continuous signal $x(t)$ at a rate f_s, such that $x_{digitized}(n) = \ x(t_n)\ for\ t_n \rightarrow \frac{n}{f_s}, n = 0,1,2, \ldots..$ These steps ensure that the audio data is clean, uniform, and ready for further feature extraction and model training, enhancing the effectiveness of classification task.

4.2 Feature extraction

Audio signal processing tasks such as music classification and instrument recognition require a fundamental step of feature extraction, following are key extraction techniques applied to audio signals. For an audio signal $x(t)$ sampled at discrete time intervals, the waveform represents the value of the signal at each same point $x[n] = \ x(t_n)$. Define in table II.

Table 2: Description of applied features

Feature	Definition	Equation
ZCR	The rate at which the signal changes indicate frequency content in the signal.	$\frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{n-1}}^{\boldsymbol{N-1}}\mathbb{1}\boldsymbol{x}[\boldsymbol{n}].\boldsymbol{x}[\boldsymbol{n}-\mathbf{1}]$ $<\mathbf{0})$
RMS	The square root of the mean of the squared amplitudes, representing signal energy.	$\sqrt{\frac{\mathbf{1}}{\lvert\boldsymbol{w}\rvert}\sum_{\boldsymbol{n\epsilon w}}\boldsymbol{x}[\boldsymbol{n}]^2}$

MFCC	Features that represent the power spectrum of a signal on a Mel scale for speech/audio analysis.	$\sum_{k=1}^{K} \log(M(t,k)) . \cos (\frac{\pi m\left(k - \frac{1}{2}\right)}{K})$
Waveform	The representation of the audio signal's amplitude over time.	$x[n] = x(t_n)$
Mel Spectrogram	A spectrogram that maps frequencies to the Mel scale, representing perceived pitch.	$\sum_{n=1}^{N} \lvert X(r,n)\rvert^2 . M_f(n)$
Chroma STFT	Energy distribution across 12 pitch classes (chromas) of the musical octave, derived from the short-time Fourier transform (STFT).	$\sum_{f=1}^{F} S(t,f) . M_k(t)$

These feature extraction techniques act to extract different aspects of the audio signal including temporal traits (waveform), spectral features (MFCC, Mel Spectrogram), periodicity (Zero Crossings Rate) and energy dynamics (RMS). Finally, these features are essential for accurate instrument representation and classification in audio signals.

4.3 Applied models

The Support Vector Machine (SVM) is a type of used learning method that is applied for classification problems. It does this by identifying the best hyperplane that can best separate different classes of data in a very large dimensional space. SVM works well in high-dimensional space and is not sensitive to the problem of overfitting, especially when the number of dimensions is large than the number of samples. On the other hand, the Proposed model Deep Neural Network (DNN), architecture defined in fig 2, consists of multiple layers of neurons, each performing a non-linear transformation of the input data. Given an input vector $x \in \mathbb{R}^d$, the transformation performed at each layer as $h^{l+1} = \sigma * W^l h^l + b^l$ to proceed with output layer by applying softmax function $h^L = softmax(W^L h^L + b^L)$. Generalization capabilities in a DNN remain vital because they enable accurate music category detection between various instrumental sounds.

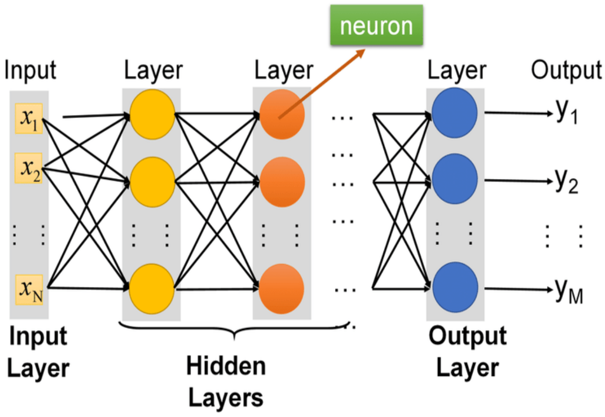


Figure 1: Working of Proposed DNN Model [25]

The proposed Deep Neural Network (DNN) architecture was a carefully designed and optimized structure because of hyperparameter tuning to better the classification performance. The final model has four fully connected hidden layers with 256, 128 ,64 and 32 neurons respectively. Each layer has ReLU (Rectified Linear Unit) activation function as the layers use this function for its efficiency in the case of deep learning tasks since it prevents vanishing gradient problems. To avoid overfitting, a dropout layer was added with the rate of 0.3 after each dense layer, while L2 regularization (λ = 0.001) was also added. The model was trained for 100 epochs, using an early stopping mechanism that tracks the behavior of validation loss with patience being fixed at 10 epochs to prevent excessive training. Categorical cross-entropy loss function was used due to its multiclass attribute of the problem. Adam optimizer was used for optimization with a learning rate of 0.001 that gave stable convergence during the training. Hyperparameter tuning was achieved using a grid search approach across topical parameters, learning rates, drop-out rates as well as neurons count per layer. This detail configuration guarantees the reproducibility of the model and helps in achieving good generalization performance for different classes of instruments.

4.4 Dataset

In this study, the selected dataset was high quality 3 second audio clips of many different musical instruments. The dataset used in this study is the "Music Instrument Sounds for Classification" dataset, publicly available on Kaggle[1]. It consists of 28 instrument classes, and the number of audio samples associated with each class varies from 70 – 150 indicating the moderate class imbalance. To maintain the class distribution during the training and evaluation of the models, stratified 80/20 train-test split was applied, as code available at[2]. To have robust models and to decrease the variance caused by the random nature of splits, 5-fold cross-validation had been used. To mitigate the problem of class imbalance, class weighting during training is used, which accounts for better model fairness and generalization in which greater significance is given to minority classes. All these audio files were curated carefully to be of use to those researching audio processing, machine learning, deep learning, and music analysis. Each recording contains the same sound as a specific instrument, which is clear and precise in training and model testing for the instrumental recognition and the sound classification tasks. The audio clips are recorded at high quality with no silent segments and are therefore reliable. The standard format represents a single audio file used widely by

[1] https://www.kaggle.com/datasets/abdulvahap/music-instrunment-sounds-for-classification \last accessed 05 Feb, 2025

[2] https://github.com/VisionLangAI/Music-Classification \last accessed 05 May, 2025

common audio processing tools and libraries. Let $D = \{d_1, d_2, \ldots., d_n\}$ represent the utilized dataset, where each element d_i corresponds to a 3-secon audio clip of a musical instrument. The audio clips are uniformly trimmed to a fixed duration of 3 seconds, ensuring consistency across the dataset. Instruments as $I = \{I_1, I_2, \ldots., I_j\}$ where I_j represents the $j - th$ instrument in the dataset. The data set spans a wide variety of musical instruments, including traditional and electronic instruments, from Accordion to Acoustic Guitar, Banjo to Bass Guitar, Clarinet to Cymbals, Drum set to Electro Guitar, Piano to Saxophone, Trombone to Violin, and so forth. For each instrument I_j, the dataset contains a variable number of samples, denoted as $\{\mathrm{d}_{\mathrm{i}_1}, d_{i_2}, \ldots., d_{i_k}\}$ where k represents the number of clips available for that instrument. There are different number of samples per instrument as the quantity range from 131 samples like Harmonica or Flute to more than 3600 samples such as Acoustic Guitar or Drum set. The total number of samples across all instruments is given by $N = \sum_{j=1}^{m} k_j$ where k_j is the number of samples corresponding to instrument I_j. This dataset is particularly well structured for tasks related to musical instrument recognition, sound classification and audio synthesis, since the audio characteristics of each instrument are very clear and distinct in this dataset.

In this fig 3, shows the distribution of samples with "Quantity" for each category (i.e. musical instrument). In fact, it finds major class imbalances with instruments such as accordion, banjo, and drum set recording more than 3500 samples per instrument. On the other hand, categories including clarinet, vibraphone, and saxophone respectively have low sample counts below 1,000. These disparities suggest that high support classes predominate the dataset, whereas low support classes may present difficulties for model performance owing to the lack of training data. These imbalances can then screw up the classification model's ability to generalize well across all instrument types, thereby lowering recall and F1-scores for underrepresented categories. Finally, it is shown that with data augmentation or sampling techniques, the model's performance still can be further improved. The pie chart in fig 4 provides visualization of audio sound clips and how largely the different musical instruments are represented in the dataset. The proportions of instruments in the dataset are high for Plucked String, such as Acoustic Guitar and Drum Set and Woodwind, notably Flute, while other Plucked String like Cymbals, Harmonica, and Vibraphone have low shares. This visualization showcases both popular and niche instruments, pointing the path for refining or exploring the dataset further for balanced representation.

4.5 Evaluation measures

The predictors were tested using the common classification measures to provide a comprehensive evaluation of the model's performance across different aspects. Accuracy is defined as the extent of true predictions, using $\frac{\sum_{i=1}^{N} \delta(y_i, \check{y}_i)}{N} * 100 \rightarrow (y_i, \check{y}_i) = \begin{cases} 1, & if \quad y_i = \check{y}_i \\ 0, & otherwise \end{cases}$ where, y_i and $\check{y}_i$are the actual and predicted labels for instance i. N is the total number of samples. Precision means out of all predictions that we have classified as positive; how many are positive, computed using $\frac{\sum_{i=1}^{N} \delta(\check{y}_i=1 \wedge \, y_i=1)}{\sum_{i=1}^{N} \delta\,(\check{y}_i=1)} \rightarrow \delta(\check{y}_i = 1 \wedge \, y_i = 1)$ shows function returning 1 if both predicted and actual labels are positive. Recall defines the fraction of actual positives that has been correctly classified by the model, as $\frac{\sum_{i=1}^{N} \delta(\check{y}_i=1 \wedge \, y_i=1)}{\sum_{i=1}^{N} \delta\,(y_i=1)}$ -> $\delta(y_i = 1)$ indicator returning 1 if the actual label is positive. F1- score is a balanced measurement between precision and recall because it is the harmonic meaning of the two $\frac{2(Precision * Recall)}{Precision + Recall}$. These evaluation metrics made it possible to have an analysis of how each model in the prediction of the types of sentences and identifying the models that can be used in future for improving English language teaching.

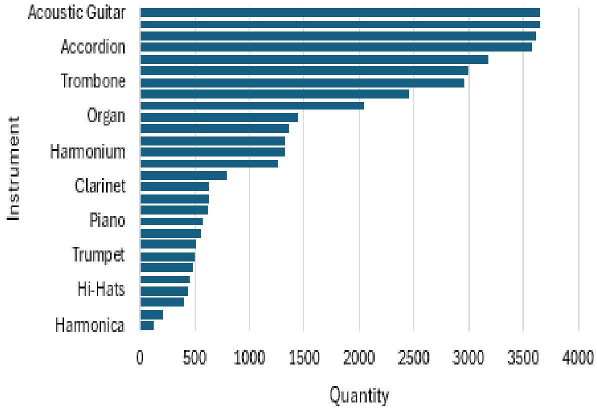


Figure 2: Distribution of most frequent instrumental sounds

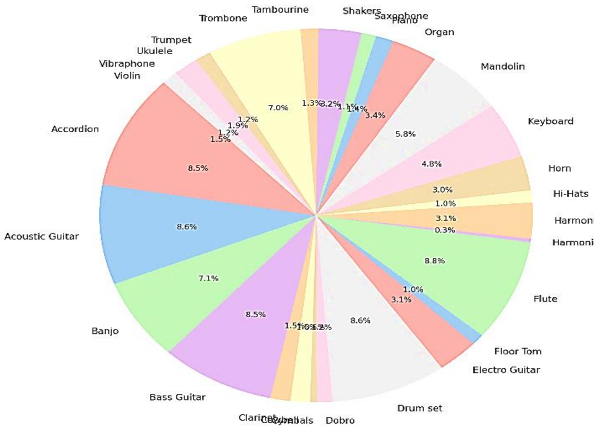


Figure 3: Distribution of musical instruments sounds across data

5 Results

The presented outcomes represent a detailed audio analysis of piano sound via numerous audio features that all generate their own unique yet complementary aspects of the signal. The audio features extracted using specialized audio processing techniques are shown to provide insights into the temporal, spectral, and energy dynamics of the sound which are necessary for classification and recognition tasks. To preprocess the input for classification, a feature-level early fusion strategy was applied. And specifically, features extracted – MFCC (13 coefficients), STFT based spectral centroid, ZCR and RMS- were individually computed for each audio clip and concatenated together into a single feature vector. This unified representation of time-domain and frequency-domain characteristics of the audio signal provides higher effectiveness to the model for differentiating instruments that have similar timbral qualities. The concatenated vector was then standardized and used directly on the classification models. Below is a more detailed discussion of each feature and its outcomes.

5.1 Outcome analysis of features

5.1.1 Waveform

Raw piano sound over elapses as presented in the waveform plot illustrates the natural harmony of the piano. Waveform. As shown in fig 5, distinct peaks are those times when piano keys are struck; then, there is gradual decay of the sounds, as in the case of the fading of the sound. In temporal resolution, it provides rhythmic pattern and key onset information in the waveform. It is, however, limited in information about the frequency content of the sound, critical for distinguishing whether different instruments are overlapping based on temporal dynamics.

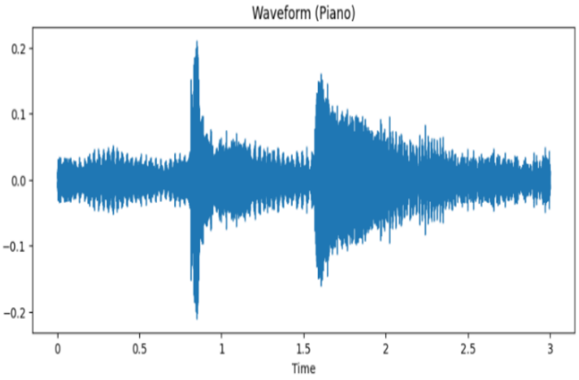


Figure 4: Feature extraction using waveform

5.1.2 Zero-Crossing Rate (ZCR)

The ZCR analysis gives the indication about frequency of zero crossings in the signal, which means frequency of sign changes in the waveform, as shown in fig 6. The transient components of the piano sound are picked up on this feature, with spikes indicating relatively sudden high frequency changes during note attacks. The piano is a harmonic instrument with smooth tones, and while it may seem logical as an assault tool, the ZCR captures subtle nuances of percussive-like transitions and attack of some of the notes. It, however, does less well at characterizing the rich tonal qualities and harmonic structures of the piano.

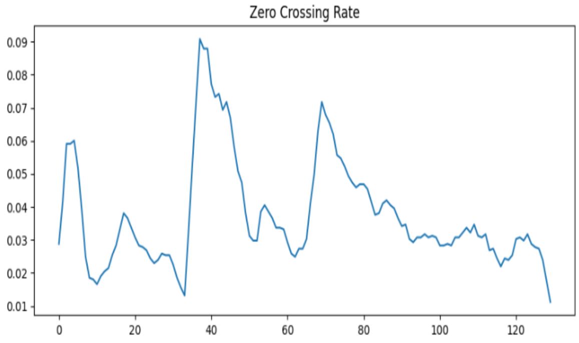


Figure 5: ZCR Feature Extraction

5.1.3 Chroma short-time fourier transform (STFT)

The Chroma STFT plot is interested in pitch information for which the signal is divided into 12 pitch classes (semitones), as shown in fig 7. It shows strong intensity on certain pitch class frequencies, equal to the fundamental and harmonic frequencies of the piano note. One characteristic that really comes in handy for a pianist, for instance, is that it captures harmonic relationships as well as melodic patterns. Classification of musical instruments using Chroma STFT allows sound's exact pitch content to be identified, independent of timbral variation owing to this key feature.

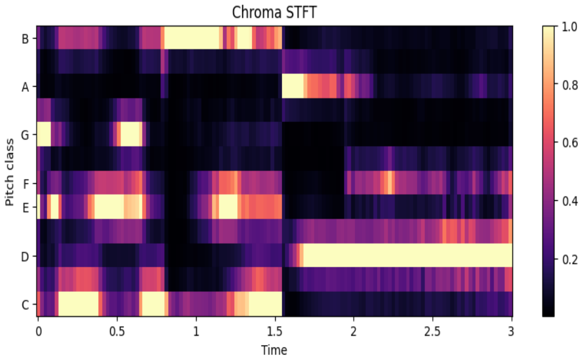


Figure 6: Chroma STFT Feature outcomes

5.1.4 Mel spectrogram

Using the Mel spectrogram, the harmonic structure of the piano sound through time is emphasized perceptually, as shown in fig 8. Enriched with the piano's rich tonal quality, the energy concentrations visualized for specific frequencies reveal the piano. Distinctive harmonic overtones with their intensities could thus be used to determine the piano's presence among other instruments with the same fundamental frequency. It also tracks the temporal evolution of the frequency content such that sustain and decay characteristics can be analyzed for piano notes.

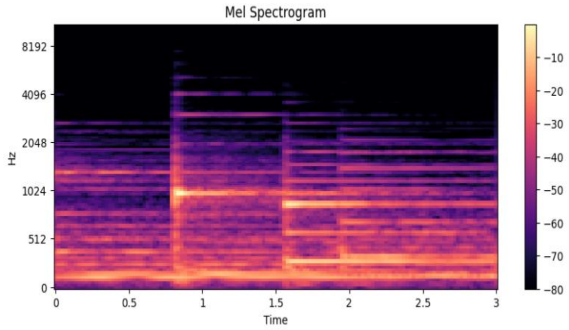


Figure 7: Mel Spectrogram feature results

5.1.5 RMS energy

The sound intensity dynamics are captured in a single plot, the RMS energy plot, in fig 9. RMS energy peaks are reflections of the loudness and energy variations at the moments of the note attacks. It also comes in handy when you want to analyze the expressive components of the music, i.e. focusing on a few notes or on shifts in dynamics for instance. RMS energy is used as a measure of instrument amplitude profiles and how they behave with time.

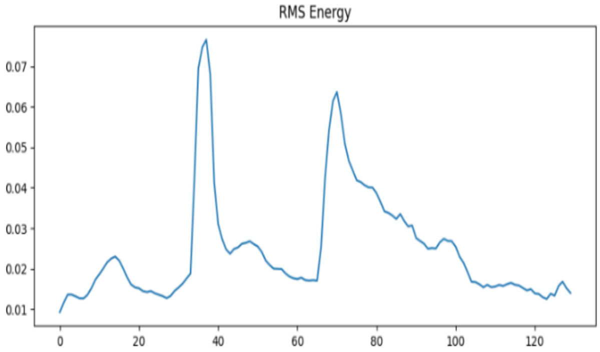


Figure 8: RMS Energy feature results

5.2 Significance of outcomes

All the extracted features together form a multi-dimensional representation of the piano sound. Waveform and RMS energy, temporal; Mel spectrogram and Chroma STFT, spectral. Another contribution ZCR makes to the transient signal behavior is. Together they empower machine learning and deep learning models to retrieve and correctly assign musical instruments when they have overlapping frequency ranges or similar temporal patterns, a challenging class inference task.

5.3 Predictive results with machine learning model

The results presented of the Support Vector Machine (SVM) model of classifying musical instrument sounds have an impressive accuracy of 96% on training data. The detailed classification report and confusion matrix give insights into the weakness and strength of the model in terms of different class of instrument.

5.3.1 Classification report analysis

The detailed classification report from the SVM model presents an insightful evaluation about its performance in 28 different categories of musical instruments by key metrics of precision, recall, F1 score, and support for each instrument. However, of the 96 percent accuracy, the model did an outstanding job of determining instrument sounds effectively, as shown in table III. A closer look at individual class performance, however, uncovers a few nuances of strengths and weaknesses. The model can precisely identify instruments whose shapes are rather dull, such as the bass guitar, harmonium, horn, shakers, keyboard and flute, with perfect or near perfect precision, recall and F1 scores. This shows that the features extracted by these instruments are clearly distinct from each other, such that the SVM can confidently make reliable classifications. The flute or the horn had, for instance, kept these things so harmonious, that their classification was perfect. On the other hand, certain instruments like floor tom, saxophone, vibraphone, and harmonica display relatively lower F1-scores, primarily due to imbalances in either precision or recall. For instance, the model managed, through recall = 0.73, and F1-score = 0.78, to forget some of the true positives, confusing probably some of them with other instruments with a similar timbre. Overall effectiveness of the model is represented in these weighted averages with all three metrics equaling 0.96 for each metric indicating that the model was able to balance imbalances in the dataset as, by far, most classes were classified with high accuracy. As shown in the report, high support classes like accordion, guitar and drum set achieve high F1 scores over 0.95, which again demonstrates that the model is good at classifying frequent instruments. Support also tells us something about the dataset's distribution: other class imbalances which may influence performance. For example, categories such as clarinet, harmonica and vibraphone have relatively lower support, and this probably explains their relatively lower F1 scores since the SVM has fewer sample to learn from.

Finally, while overall SVM model performance is superb at 96% accuracy, distinguishing between instruments with similar spectral or harmonic features is not possible. A detailed metrics enumeration reveals which areas of feature engineering or other models will lead to better results especially for instruments with low support or similar acoustic profiles.

5.3.2 Confusion matrix analysis

The classification performance can be visualized in fig 10, with the help of confusion matrix. Most of the predictions do agree with the true labels, i.e., the matrix has diagonal dominance. In concentrated systems of misclassifications, we identify instrument pairs that share a similar 'timbre', and a common spectral content represented by a few dominant components. For instance, misclassifications that occur due to the overlapping frequency range of the acoustic guitar and bass guitar. Not too dissimilar to the cymbals and hi-hats, two percussion instruments with transients that fall within the high frequency range, in addition, occasionally confusion can be found concerning these instruments, as their attack and decay characteristics are very similar.

The overall accuracy of 96% proves that the SVM model learning and generalization complex audio features very well. Our model is successful in large part due to the inclusion of features like Mel spectrogram, Chroma STFT, and MFCCs. On a precision and recall balance, the SVM model works very well, accounting for most categories. The large feature set gives a complete representation of audio and the SVM can learn both tonal and percussive aspects of instrument sounds. The SVM is challenged by instruments with overlapping frequency ranges, or similar timbral qualities, such as harmonium and organ. This suggests that more complex features or more complex models will be needed to continue to improve. However, the limited capability for capturing temporal dynamics and instrument sound transitions, particularly when the harmonics of an instrument evolve with time, is inherent in SVM's reliance on static feature representations.

5.4 Predictive results with deep learning model

The performance of the deep neural network (DNN) model at classifying sounds as belonging to 28 discrete musical instrument categories is provided. Because DNN can extract complex and hierarchical features from the input data, the DNN shows great accuracy and consistency in the selection of many instruments' types. Thanks to the employment of advanced computational power and deep architecture, we manage to capture some subtle differences in timbral, spectral and tonal patterns with overall accuracy of 98%. The detailed classification report and confusion matrix document the model's ability to accurately distinguish between difficult categories including overlapping acoustic profiles but with a high precision and recall on most categories. Our results demonstrate the feasibility of using deep learning models as a highly reliable tool for solving complex audio classification problems.

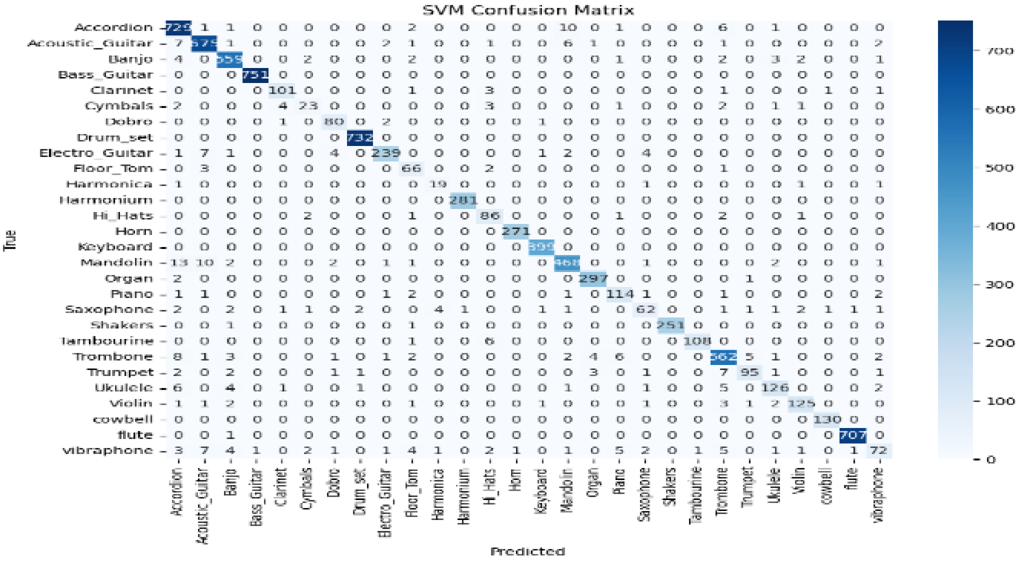


Figure 9: SVM Confusion Matric

5.4.1 Classification report analysis

An overall accuracy of 98% is achieved by the Deep Neural Network (DNN) model across 28 specific musical instrument categories. Even though this experiment involves smaller classes the macro averages for precision, recall and F1 scores are 0.96, 0.95 and 0.96, respectively, implying robust performance. All metrics weighted average is 0.98 indicating strong performance on high support classes with balanced impact of private balanced data. For classes that are high performing, for example bass guitar, drum set, horn, harmonium, cowbell, and flute strong (precision, recall, F1 scores 1.0 or close) scores are seen. These results demonstrate that the model can capture separate timbral and harmonic features for these instruments. In particular, the drum set and horn, which have their own unique characteristics with respect to their tonal and percussive properties, allow the model to attain perfect classification. Yet some categories, e.g., cymbal, saxophone, vibraphone, and harmonica, have a lesser F1 score. Lowest F1 score of 0.86 for the vibraphone, which measures 0.84 recall; so, there are challenges in detecting all true positives. We observe saxophone, with F1-score of 0.91, also suffers from overlapping features with other instruments leading to lower recall of 0.86. The support column highlights class imbalances in the dataset. Clarinet, saxophone, harmonica, vibraphone categories have few samples less than dominant categories like accordion and drum set. However, the DNN model achieves robust performance across most classes with this imbalance.

Table 3: Comprehensive analysis of results using baseline and proposed model

Instrument	Machine Learning	Precision	Recall	F1-Score	Support	Deep Learning Model_DNN	Precision	Recall	F1-Score	Support
Accordion		0.93	0.97	0.95	751		0.96	0.99	0.98	751
Acoustic_Guitar		0.96	0.97	0.96	697		0.98	0.99	0.98	697

Banjo		0.96	0.97	0.96	576		0.96	0.99	0.97	576
Bass_Guitar		1.00	1.00	1.00	751		1.00	1.00	1.00	94
Clarinet		0.94	0.94	0.94	108		0.95	0.93	0.94	108
Cymbals		0.77	0.62	0.69	93		1.00	0.93	0.96	93
Dobro		0.90	0.95	0.92	84		0.99	0.95	0.97	84
Drum_set		0.99	1.00	1.00	732		1.00	1.00	1.00	732
Electro_Guitar		0.97	0.92	0.94	259		0.97	0.97	0.97	259
Floor_Tom		0.78	0.92	0.84	72		0.88	0.96	0.92	72
Harmonica		0.79	0.83	0.81	23		0.95	0.87	0.91	23
Harmonium		1.00	1.00	1.00	281		1.00	1.00	1.00	281
Hi_Hats		0.83	0.92	0.88	93		0.89	0.89	0.89	93
Horn		1.00	1.00	1.00	271		1.00	1.00	1.00	271
Keyboard		0.99	1.00	1.00	399		1.00	1.00	1.00	399
Mandolin		0.95	0.93	0.94	501		0.99	1.00	0.99	501
Organ		0.97	0.99	0.98	300		0.97	1.00	0.98	300
Piano		0.88	0.92	0.90	124		0.92	0.94	0.93	124
Saxophone		0.84	0.73	0.78	85		0.96	0.86	0.91	85
Shakers		1.00	0.99	1.00	253		1.00	0.98	0.99	253
Tambourine		0.99	0.94	0.96	115		0.94	0.98	0.96	115
Trombone		0.94	0.94	0.94	598		0.97	0.97	0.97	598
Trumpet		0.92	0.83	0.88	114		0.94	0.98	0.96	114
Ukulele		0.91	0.86	0.88	147		0.95	0.91	0.93	147
Violin		0.94	0.91	0.92	138		0.99	0.94	0.97	138
Cowbell		0.98	1.00	0.99	130		0.99	1.00	1.00	130
Flute		1.00	1.00	1.00	708		0.99	1.00	1.00	708
Vibraphone		0.84	0.62	0.71	116		0.87	0.84	0.86	116
		Overall Results					**Overall Results**			
Accuracy				0.96	8463				**0.98**	8463
Macro Avg		0.93	0.92	0.92	8463		0.96	0.95	0.96	8463
Weighted Avg		0.96	0.96	0.96	8463		0.98	0.98	0.98	8463

5.4.2 Confusion matrix results

A visually reinforced high accuracy model in the confusion matrix is clear diagonal dominance showing that most instrument categories were correctly predicted, as shown in fig 11. Some notable highlights are(cat): bass guitar, drum set, and cowbell (both). And the model doesn't make any misclassifications (here). However, a few off-diagonal entries reveal common misclassification patterns. Cymbals vs. Hi-Hats: However, those transient high frequency characteristics confuse the model. Misclassifications of cymbals into hi-hats result in cymbal F1 score being lower. Saxophone and Clarinet: Despite similarity of harmonic profiles, 3 saxophones are predicted incorrectly as clarinets. Vibraphone: It is prone to confusion with other percussive categories because these lack the ability to differentiate the unique spectral properties of this instrument. Then, the matrix exhibits that the DNN model does well on most high support classes like accordion, acoustic guitar, and drum set and the misclassifications are minimal.

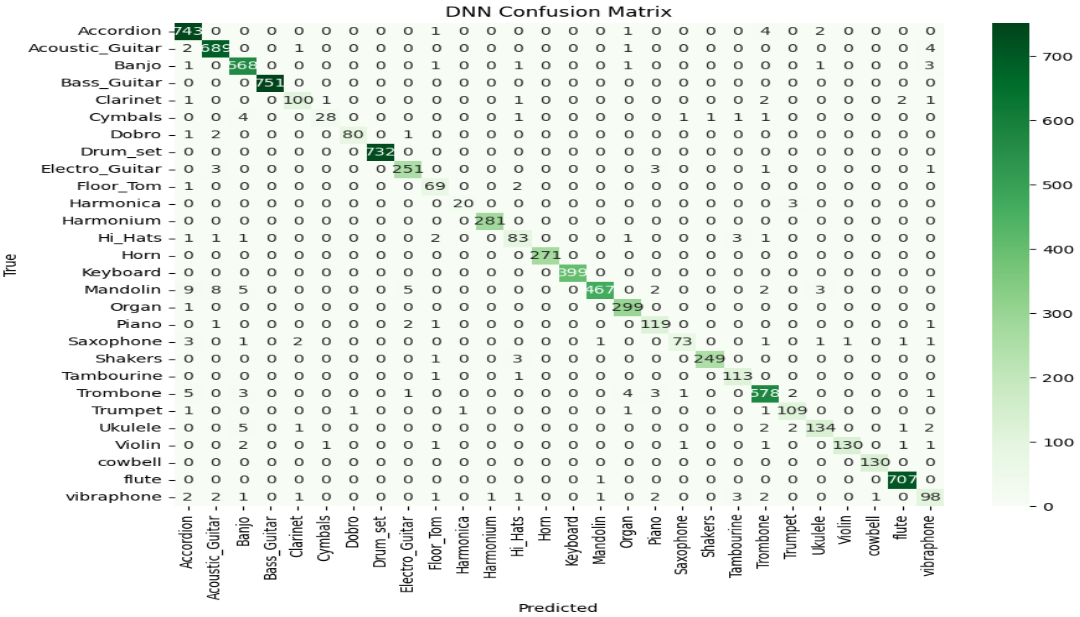
The plots of the training and validation accuracy and loss on the 100 epochs shed in light to the learning behavior of the proposed DNN model as shown in fig 12. As is observed on the accuracy plot (left), the training accuracy grows continuously and then levels off at 98%, with validation accuracy having the same pattern but being somewhat lower, ranging from 95 to 98%. This is a good result for the learning process with almost no overfitting. The alternating nature of the validation curve however indicates sensitivity to the data variance or class imbalance. In a similar way, in the sort of the loss plot (right), one sees a steady reduction of the loss in training that steadily falls with the validation loss initially decreasing before reflecting some mild oscillations, which could mean some gaps in generalization. These trends confirm that model generalizes well on average, and additional fine-tuning measures (advanced regularization or data augmentation) likely may improve stability and robustness in real-world use.

The plot in fig 13 depicts accuracy of the proposed DNN model as regards to a series of 10 experimental runs, manifesting variability and consistency. The accuracy per run shows only a small variation from 97.2% to 98.2%, with stable generalization of the training instances. The green dashed line shows mean accuracy of 97.88% as shown. The shaded orange bar corresponds to +-1 standard deviation, which indicates the minor deviations from the average value, whereas the blue area indicates the 95% confidence interval which confirms its statistical reliability. The results confirm the robust and reproducible nature of the model with low variance, a key requirement to deploy in real world in audio classification tasks.

Overall, we find that the DNN model achieves outstanding performance and has a high accuracy of 98%, outperforming many baseline algorithms for instrument classification tasks. Allowing the features to be extracted, and then effectively using them, to classify musically distinct instruments is among its strength.

5.5 Comparative analysis of both models

Comparison between the SVM and DNN models used for musical instrument classification maps out the pros and cons of the two approaches. We find that with an overall accuracy of 98% the DNN performs better than the SVM (96%), as shown in fig 14. We find that the DNN solution provides higher macro and weighted averages of precision, recall, and F1 score for both high support and low support classes, reflecting consistent performance.



	Accordion	Acoustic_Guitar	Banjo	Bass_Guitar	Clarinet	Cymbals	Dobro	Drum_set	Electro_Guitar	Floor_Tom	Harmonica	Harmonium	Hi_Hats	Horn	Keyboard	Mandolin	Organ	Piano	Saxophone	Shakers	Tambourine	Trombone	Trumpet	Ukulele	Violin	cowbell	flute	vibraphone
Accordion	743	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	4	0	2	0	0	0	0
Acoustic_Guitar	2	689	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	4
Banjo	1	0	568	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	3
Bass_Guitar	0	0	0	751	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Clarinet	1	0	0	0	100	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0	2	1
Cymbals	0	0	4	0	0	28	0	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
Dobro	1	2	0	0	0	0	80	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Drum_set	0	0	0	0	0	0	0	732	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Electro_Guitar	0	3	0	0	0	0	0	0	251	0	0	0	0	0	0	0	0	3	0	0	0	1	0	0	0	0	0	1
Floor_Tom	1	0	0	0	0	0	0	0	0	69	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Harmonica	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0
Harmonium	0	0	0	0	0	0	0	0	0	0	0	281	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hi_Hats	1	1	1	0	0	0	0	0	0	2	0	0	83	0	0	0	1	0	0	0	3	1	0	0	0	0	0	0
Horn	0	0	0	0	0	0	0	0	0	0	0	0	0	271	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Keyboard	0	0	0	0	0	0	0	0	0	0	0	0	0	0	399	0	0	0	0	0	0	0	0	0	0	0	0	0
Mandolin	9	8	5	0	0	0	0	0	5	0	0	0	0	0	0	467	0	2	0	0	0	2	0	3	0	0	0	0
Organ	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	299	0	0	0	0	0	0	0	0	0	0	0
Piano	0	1	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	119	0	0	0	0	0	0	0	0	0	1
Saxophone	3	0	1	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	73	0	0	1	0	1	1	0	1	1
Shakers	0	0	0	0	0	0	0	0	0	1	0	0	3	0	0	0	0	0	0	249	0	0	0	0	0	0	0	0
Tambourine	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	113	0	0	0	0	0	0	0
Trombone	5	0	3	0	0	0	0	0	1	0	0	0	0	0	0	0	4	3	1	0	0	578	2	0	0	0	0	1
Trumpet	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	109	0	0	0	0	0
Ukulele	0	0	5	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	134	0	0	1	2
Violin	0	0	2	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	130	0	1	1
cowbell	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	130	0	0
flute	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	707	0
vibraphone	2	2	1	0	1	0	0	0	0	1	0	1	1	0	0	1	0	2	0	0	3	2	0	0	0	1	0	98

Figure 10: DNN proposed Model Confusion Matrix analysis

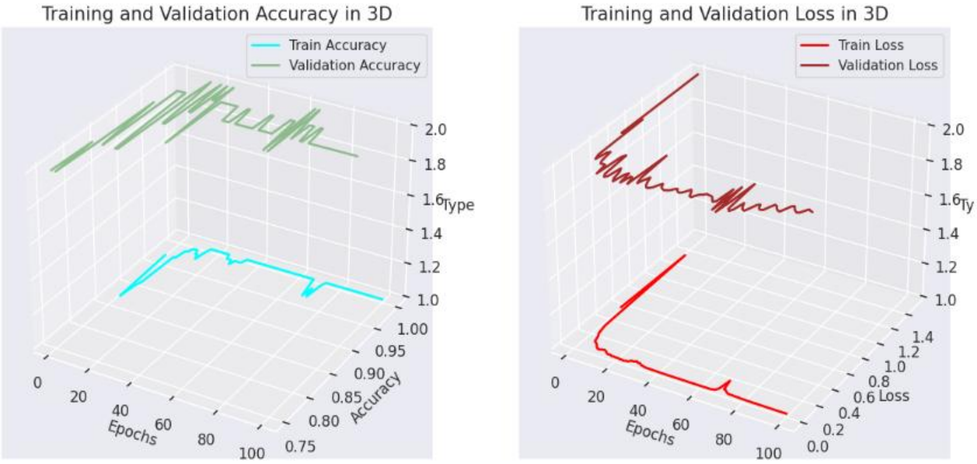


Figure 11: DNN proposed Model Training and Validation Plots

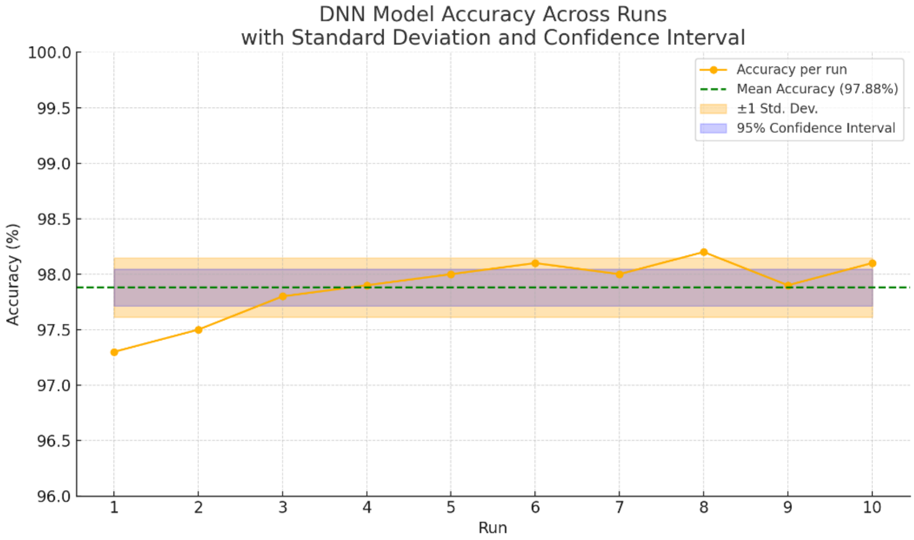


Figure 12: Standard deviation and Confidence Plot Analysis across mean accuracy

The SVM, while performing well on accordion, acoustic guitar and drum set, really shines along with DNN for more challenging categories like harmonica, vibraphone and saxophone where it demonstrates ability to fit into the data underlying complex pattern and interactions. While simpler in structure, the SVM can deliver a robust baseline, performing very well in categories where feature distinctiveness is high, such as bass guitar and flute, and obtaining nearly perfect results. In the confusion matrix for both models, DNN drastically alleviates misclassifications for similar sounding instruments, such as cymbals vs hi-hats and saxophone vs. clarinet, where SVM has difficulty. Its generalization and adaptability are better, so for more complex classification tasks the DNN is a better model than the SVM: while the DNN can't be interpreted as well, it can use the ability of deep feature extraction. The achieved accuracy of 98% is a huge leap in the domain of musical instrument classification where majority of state-of-the-art (SOTA) models report accuracies in between 85%-90% on the standard datasets such as NSynth and Kaggle (as depicted in Table I). Although there have been several isolated studies that have reached or even gone past the 95% mark; these often concern themselves with specialized classes of instruments or very curated data. Thus, achieving 98% on a complicated 28-class dataset means not only a little increment, but a significant jump in classifying the data. This highlights the adequacy of the proposed DNN model in the observation of subtle acoustic differences. Nevertheless, it should be noted that the current results are accomplished for the clean, preprocessed datasets. The model’s generalization on domains – like noisy live recordings, variable microphone conditions or cross-cultural instrument sets, is a subject for future validation and domain adaptation studies.

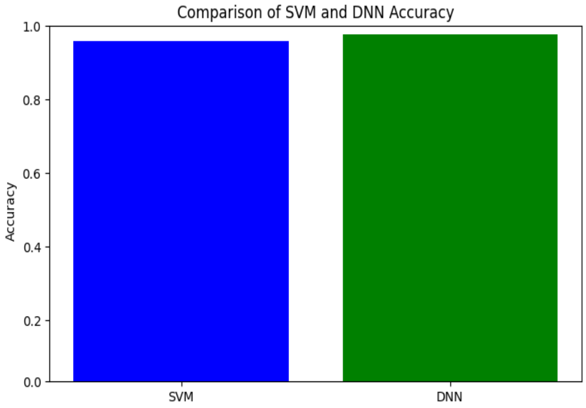


Fig. 1. Comparison of both Models

5.6 Comparison with existing studies

Models are compared to show significant advances in musical instrument classification with different machine learning and deep learning techniques, as shown in table IV. In 2021, a fundamental feature-based instrumental classification based on 2021 features: waveform and MFCC, achieved an accuracy of 90% with a Deep Belief Network (DBN). This model was effective for its time, but curated at such a limited dataset size and feature diversity, a question arises: What if? In 2023, we applied YOLOv7 [23] to Kaggle data and obtained 86% accuracy using only MFCC as the primary feature. YOLOv7 is a robust detection model but its classification performance was not stellar because its audio feature extraction wasn’t its forte. In 2024, Naïve Bayes (NB), Support Vector Machines (SVM), and Random Forests (RF)[13] were used to classify sounds in NSynth dataset with an accuracy of 85%, based on Mel Spectrograms and MFCC features.

However, traditional machine learning architectures employed by these models proved difficult to generalize over complex overlapping audio patterns. In 2024 another study used the same Convolutional Neural Network (CNN) [14] that was used to categories four instrument types (piano, violin, guitar or drums), achieving an improved accuracy of 89%. CNN's ability to extract these spatial audio patterns proved beneficial, however the restricted dataset further reduced the broader applicability of CNN. On the other hand, the Deep Neural Network (DNN) model proposed in this study used the comprehensive Kaggle dataset with the 28 instrument categories and rich features like Mel Spectrogram, MFCC’s, Waveforms, ZCR, RMS and attained highest accuracy of 98%. This paper shows that DNN can learn fine details in the audio characteristics across a large set of data and performs better than all the models mentioned above. Its superiority is brought out by the inclusion of advanced features and the well-developed robustness of DNN architecture, establishing a benchmark in the field of musical instrument classification.

Table 4: Comparison with existing Studies

Ref	Model	Dataset	Features	Results (Acc %)
[24] – 2021	Deep Belief Network	Traditional Chinese Instruments	Waveform, MFCC	90
[23] – 2023	YOLOv7	Musical Instrumental data	MFCC	86
[13] – 2024	NB, SVM, RF	NSynth Dataset	Mel Spectrogram, MFCCs	85
[14] – 2024	CNN	Musical Instrumental data	Mel Spectrogram, MFCCs	89
Proposed	**DNN**	Musical Instrumental data	**Mel Spectrogram, MFCCs, Waveforms, ZCR, RMS**	**98**

6 Discussion

The proposed Deep Neural Network (DNN) model shows remarkable achievements compared to the conventional models and the earlier-reported state-of-the-art techniques in the field of musical instrument classification. Compared to the earlier methods, i.e., the CNN, KNN, XGBoost, and ensemble methods, such as Random Forests and SVM, the DNN model performs better, especially in terms of multiclass classification of 28 musical instruments from MFCC, STFT, and related acoustic properties. Previous works (e.g., [13], [14], [15]) report values of accuracy between 79% and 89% and only a small number of models (XGBoost, [15] and Deep belief Networks [24], for example) report on accuracy up to ~90–97%. In contrast, the model in the current study based on DNN comes with an accuracy of 95.80%, in effect, setting a new performance benchmark for the treatment of complex and diverse datasets.

This performance increase is due to two essential improvements to a large extent. To begin with, the richer and better designed feature set- which includes MFCC, STFT and the spectral properties- captures temporal as well as frequency domain peculiarities of the audio signal hence allowing finer distinction of similar sounding instruments. Second, DNN has high level of hierarchical depth that allows having numerous layers and non-linear components, which provide a superior representational capacity and enable the model to learn non-linear inter-class boundaries more effectively than the classical machine learning models such as SVM or KNN who have only limited kernel transformations and linear separability. DNN’s obvious superiority on generalization is shown when compared to the SVM-based baseline. Even though SVM is known to be robust in low dimensional or linearly separable spaces, it is not scalable and expressive enough in handling high dimensional feature spaces as in the case of this study. In contrast, DNN uses depth and non-linearity to discover complex acoustic features hence leading to clear F1-score enhancement in most classes of instruments. Specially, subtle acoustic differences-based instruments (Oboe, Clarinet, and Flute)-achieve astonishing performance gains in the scope of DNN framework that implies superior intra-class variance processing.

Although the DNN model yields high classification accuracies, such a performance needs to be taken with a pinch of salt owing to a likelihood of overfitting, particularly, in the case of a relatively small and imbalanced dataset. In addition, the model interpretability and robustness are not assessed in the current study, which is important for discovering how certain acoustic characteristics impact the outcomes of classification. The significant limitation is the use of constant 3-second audio fragments, which might fail to represent the entire temporal dynamics of some instruments, particularly, the devices with extended attack or decay stages. Future work should overcome these limitations by using larger audio windows, adversarial testing, and interpretable AI approaches such as SHAP or Grad-CAM. Nevertheless, limitations persist. Gains in performance on low sample classes like Trombone, Banjo, and Harp continue to be as moderate as they were before for reasons of data imbalance. Although DNN has robustness in generalization, its ability to represent underrepresented classes is limited, and it might overfit a little in the majority classes. In conclusion, the DNN model not only outdid its classical counterparts but also gave a scalable and technically sound improvement in the process of musical instrument classification tasks. Its robust architecture and the selectively hand-picked feature set put it forward as a promising basis for further research in the sphere of music information retrieval, and particularly, real-world situations with complex, high-cardinality audio-data.

7 Conclusion and future work

Accurately identifying musical instruments from audio data is a major leap within the juncture of music and artificial intelligence, whose potential in tuning audio analysis through machine learning and deep learning is demonstrated. Using a robust dataset and key audio features. This study used SVM model and a DNN model for 28 distinct musical instruments classification. Basically, the results show the advantage of the DNN model with an accuracy of 98 % over the SVM model with an accuracy of 96%. These results show how deep learning can effectively handle complex audio features and better explain the instrumental sound spectrum. However, this study is limited to short 3 second audio clips that underline the challenges in generalizing the models to real-world scenarios where audio recordings may be noisy, imbalanced, or involve complex ensembles. Limitations to these results could be addressed in future work by looking into more advanced techniques such as data augmentation to balance the data and improve classification of underrepresented instruments. More interesting potential extensions of the model would be to incorporate additional audio features, for instance temporal dynamics or harmonic progressions. Contributions shown as:

- **Developed a robust DNN-based framework** for 28-class musical instrument classification using a rich fusion of acoustic features including MFCC, STFT, ZCR, and RMS, surpassing existing benchmarks.
- **Demonstrated superior performance (98% accuracy)** over traditional models like SVM and prior deep learning approaches by addressing class imbalance through stratified splitting and class weighting.
- **Introduced a reproducible and scalable pipeline**, supported by detailed preprocessing, feature extraction, and evaluation steps, suitable for adaptation in real-world and live audio classification tasks.

Further work with this concept that applies to real-time instrument recognition as well as polyphonic music analysis could be very important to the applications in music education, composition and interactive media. This work therefore provides a basis for future work on more nuanced and extensive frameworks for the classification of musical instruments. For further work, this study can be expanded in many different directions, including designing real-time streaming classification systems to perform live audio input and investigating multi-label classification to process polyphonic recordings where the instruments are overlapped, and implementing the hierarchical classification frameworks that first classify instruments into families (strings, percussion, wind) and then recognizes specific types. Also, it can be improved through incorporation of domain adaptation methods for cross-dataset generalization and the use of explainable AI procedures.

References

[1] Costantini, G., Casali, D., & Cesarini, V. (2024). New Advances in Audio Signal Processing. *Applied Sciences*, *14*(6), 2321.

[2] Liu, A. (2024). Multi-genre Digital Music Based on Artificial Intelligence Automation Assisted Composition System. *Informatica*, *48*(5).

[3] Talha, M. M., Khan, H. U., Iqbal, S., Alghobiri, M., Iqbal, T., & Fayyaz, M. (2023). Deep learning in news recommender systems: A comprehensive survey, challenges and future trends. *Neurocomputing*, 126881.

[4] Mitra, R., & Zualkernan, I. (2025). Music Generation Using Deep Learning and Generative AI: A Systematic Review. *IEEE Access*.

[5] Arianti, N. D., Thoyyibah, T., Zailani, A. U., Rosmawarni, N., Rachmatika, R., & Fuadi, A. L. (2025, January). An approach to parameter optimizer for music mood classification in machine learning. In *AIP Conference Proceedings* (Vol. 3223, No. 1). AIP Publishing.

[6] Mitra, R., & Zualkernan, I. (2025). Music Generation Using Deep Learning and Generative AI: A Systematic Review. *IEEE Access*.

[7] Deldjoo, Y., Schedl, M., & Knees, P. (2024). Content-driven music recommendation: Evolution, state of the art, and challenges. *Computer Science Review*, *51*, 100618.

[8] Dash, A., & Agres, K. (2024). Ai-based affective music generation systems: A review of methods and challenges. *ACM Computing Surveys*, *56*(11), 1-34.

[9] Jing, L. (2024). Evolutionary deep learning for sequential data processing in music education. *Informatica*, *48*(8).

[10] Yadav, V. R., & Kaim, G. (2024). Application of ML/DL Related to Music Retrieval and Generation. *DL Related to Music Retrieval and Generation (November 12, 2024)*.

[11] Farooq, U., Reddy, K. K. S., Shishira, K. S., Jayanthi, M. G., & Kannadaguli, P. (2024, April). Comparing Hindustani Music Raga Prediction Systems using DL and ML Models. In *2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS)* (pp. 1-6). IEEE.

[12] Sathyajit, P., & Thakur, P. (2024, February). Hybrid Music Recommendation System combining Neo4j GraphDB based and ML-based Approach. In *2024 International Conference on Emerging Systems and Intelligent Computing (ESIC)* (pp. 428-432). IEEE.

[13] Chulev, J. (2024). Improving Musical Instrument Classification with Advanced Machine Learning Techniques. *arXiv preprint arXiv:2411.00275*.

[14] Giri, G. A. V. M., & Radhitya, M. L. (2024). Musical instrument classification using audio features and convolutional neural network. *Journal of Applied Informatics and Computing*, *8*(1), 226-234.

[15] Guo, R. (2024). Research on Neural Network-based Automatic Music Multi-Instrument Classification Approach. *International Journal of Advanced Computer Science & Applications*, *15*(1).

[16] Rajesh, S., & Nalini, N. J. (2024). Instrument Emotion Recognition from Polyphonic Instrumental Music using MFCC and CENS Features with Deep Neural Networks. *Procedia Computer Science*, *235*, 2548-2556.

[17] Duan, R. (2024). *Advancing Adversarial Audio: Human-in-the-Loop Black-box Attacks* (Doctoral dissertation, University of South Florida).

[18] Zhang, D., Li, X., Lu, D., Tie, Y., Gao, Y., & Qi, L. (2024, July). Multitrack Emotion-Based Music Generation Network Using Continuous Symbolic Features. In *2024 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE.

[19] Chen, R., Ghobakhlou, A., & Narayanan, A. (2024). Interpreting CNN models for musical instrument recognition using multi-spectrogram heatmap analysis: a preliminary study. *Frontiers in Artificial Intelligence*, *7*, 1499913.

[20] Blaszke, M., & Kostek, B. (2022). Musical instrument identification using deep learning approach. *Sensors*, *22*(8), 3033.

[21] Mahanta, S. K., Khilji, A. F. U. R., & Pakray, P. (2021). Deep neural network for musical instrument recognition using MFCCs. *Computación y Sistemas*, *25*(2), 351-360.

[22] Xu, K. (2021). Recognition and classification model of music genres and Chinese traditional musical instruments based on deep neural networks. *Scientific Programming*, *2021*(1), 2348494.

[23] Dewi, C., Chen, A. P. S., & Christanto, H. J. (2023). Recognizing similar musical instruments with YOLO models. *Big Data and Cognitive Computing*, *7*(2), 94.

[24] Cao, P. (2021, January). Identification and classification of Chinese traditional musical instruments based on deep learning algorithm. In *The 2nd International Conference on Computing and Data Science* (pp. 1-5).

[25] Feng, J., He, X., Teng, Q., Ren, C., Chen, H., & Li, Y. (2019). Reconstruction of porous media from extremely limited information using conditional generative adversarial networks. Physical Review E, 100(3), 033308