**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | Capstone Project |
| **Assessment Title:** | Using Deep Learning to Classify and Analyse Musical Instruments Based on Spectrograms and Audio Features |
| **Supervisor Name:** | David McQuaid |
| **Student Full Name:** | Obinna Igbodika |
| **Student Number:** | 2022551 |
| **Assessment Due Date:** | 26/05/2023 |
| **Date of Submission:** | 28/05/2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Contents**

[**1. INTRODUCTION 1**](#_Toc146214617)

[**2. Research Question 2**](#_Toc146214618)

[**3. Objectives 2**](#_Toc146214619)

[**4. RELATED WORK 2**](#_Toc146214620)

[**4.1 Exploring MFCC for Robust Instrument Classification 2**](#_Toc146214621)

[**4.2 Classification Using Deep Neural Networks 3**](#_Toc146214622)

[**5. VALIDITY 10**](#_Toc146214623)

[**6. Sampling Strategy 11**](#_Toc146214624)

[**7. Primary Research Methodology 11**](#_Toc146214625)

[**8. Ethics 12**](#_Toc146214626)

[**9. METHODOLOGY 12**](#_Toc146214627)

[**10. DATASET 13**](#_Toc146214628)

[**11. FEATURE EXTRACTION 14**](#_Toc146214629)

[**12. Model Architecture 15**](#_Toc146214630)

[**13. IMPLEMENTATION 18**](#_Toc146214631)

[**13.1 SVM DECISION TREES AND RANDOM FORESTS 19**](#_Toc146214632)

[**13.2 NEURAL NETWORKS 22**](#_Toc146214633)

[**14. Discussion 25**](#_Toc146214634)

[**15. CONCLUSION 30**](#_Toc146214635)

[**16. Future Research Directions 31**](#_Toc146214636)

**Using Deep Learning To Classify And Analyze Musical Instruments Based On Spectrograms And Audio Features**

# INTRODUCTION

The realm of music and audio analysis has witnessed remarkable progress in recent years, thanks to the advent of machine learning and deep learning techniques. This progress has not only enhanced our understanding of music but also paved the way for innovative applications in music classification and recognition. In particular, the ability to classify various musical instruments within audio recordings has been a focal point of interest.

This introduction sets the stage for a comprehensive exploration of the methodologies and findings related to the classification of musical instruments based on spectrograms and audio features using a combination of machine learning and deep learning models.

The study revolves around the central objective of developing accurate and effective models for the classification of musical instruments. This entails extracting meaningful information from audio recordings, transforming it into usable data, and employing advanced algorithms to make informed instrument classifications. Such capabilities hold tremendous promise for music information retrieval, musical composition, and even music education.

To address this challenge, the study investigates a range of methodologies and models, each with its unique strengths and weaknesses. These include the utilization of spectrograms and Mel-frequency cepstral coefficients (MFCCs) as fundamental audio representations. Additionally, it explores the integration of deep learning models like convolutional neural networks (CNNs) and artificial neural networks (ANNs), alongside traditional machine learning techniques such as support vector machines (SVMs), decision trees, and random forests.

The study emphasizes the significance of accurate musical instrument classification, not only as a tool for musicologists and audio engineers but also as an enabler of innovative applications in fields like music recommendation systems, sound synthesis, and even human-computer interaction.

Throughout the subsequent sections, we delve into the intricate details of data preparation, feature extraction, model development, and rigorous evaluation. The study aims to shed light on which combinations of techniques and models offer the highest accuracy and effectiveness in classifying musical instruments. In doing so, it contributes to the broader discourse on the fusion of music and technology, with practical implications for both musicians and technologists alike.

# Research Question

What combination of spectrogram and audio features, along with the choice of deep learning and machine learning model, leads to the highest accuracy in classifying various musical instruments based on spectrograms and audio features?"

# Objectives

* To develop a deep learning model capable of accurately classifying different musical instruments based on their spectrograms and audio features.
* To investigate which combination of spectrogram and audio features yields the best classification accuracy for different musical instruments.
* To compare the performance of different deep learning models, such as convolutional neural networks and artificial neural networks, in classifying musical instruments based on spectrograms and audio features

# RELATED WORK

## Exploring MFCC for Robust Instrument Classification

Deng, Simmermacher, & Cranefield, (n.d) presents an empirical study on feature analysis for classical instrument recognition using machine learning techniques. The authors analyze three major feature extraction schemes and evaluate their performance using classifiers undergoing cross-validation. The results reveal significant redundancy between and within feature schemes commonly used in practice. The study shows that the MFCC feature scheme gives the best classification performance and that some MPEG-7 features are not reliable for robust classification results. The authors suggest further feature analysis research to optimize feature selection and achieve better results for the instrument recognition problem. They propose investigating new feature schemes and mechanisms to combine feature schemes to improve classification performance. The paper highlights the importance of feature analysis and selection for instrument recognition tasks and provides relevant information for future research in this area. Overall, the study contributes to advancing the understanding of feature analysis and selection for classical instrument recognition, and the findings have implications for improving the accuracy of machine learning-based classification systems. Overall, this study provides relevant information on the importance of feature analysis and selection for instrument recognition tasks, highlighting the need for further research in this area. The findings suggest that the MFCC feature scheme is most effective for classification tasks and that the current feature schemes used in practice are highly redundant. This paper also identifies possible areas for improvement, such as investigating new feature schemes and mechanisms to combine feature schemes to improve classification performance.

Interdisciplinary research conducted by Racharla et al., (2020), focused on retrieving information from music. The IRMAS (Instrument Recognition in Musical Audio Signals) dataset by (Juan J. Bosch et al., 2012), has been used extensively for this purpose, as it includes a wide variety of musical clips recorded from various sources over the last century, with varying audio quality. Previous research has employed various supervised learning algorithms for this classification task, with SVM (Support Vector Machine) classifiers outperforming other state-of-the-art models with an accuracy of 79% Racharla et al., (2020). Unsupervised techniques such as Hierarchical Clustering have also shown promising results. Evaluation metrics used to assess the performance of the models include Precision, Recall, F1 score, Accuracy, and Confusion Matrix (Vakili, Ghamsari, & Rezaei, 2020). There is scope for future research in this field, including using the same approach on a different dataset, exploring the idea of classifying Indian instruments, and studying and extracting more features using signal processing techniques to improve the accuracy of instrument classification. Overall, the development of robust MIR systems will contribute to a myriad of applications, including Recommender systems, Genre Identification, and Catalogue Creation, making the entire music catalogue manageable and accessible with ease (Racharla et al., 2020).

The classification of musical instruments is a complex task within the domain of music data analysis. This study addresses the need for automating this process by leveraging machine learning techniques and acoustic feature extraction (Prabavathy, Rathikarani, & Dhanalakshmi, 2020). The objective is to accurately classify musical instruments based on features obtained from diverse instruments using contemporary algorithms. This study also aims to compare the performance of SVM and kNN in achieving this goal

The study highlights the use of Mel-frequency Cepstral Coefficients (MFCC) as a feature extraction method. Additionally, the review mentions the utilization of Sonogram-based features, providing a comprehensive representation of musical instrument sounds. The combination of MFCC with Sonogram features is also explored as a feature extraction approach.

The research collects 1284 music samples from various musical instrument databases for classification purposes. Musical instruments from different categories, such as string, woodwind, brass, and keyboard instruments, are considered. The study employs the combination of MFCC and Sonogram with SVM, as well as Sonogram with kNN, to assess the accuracy of musical instrument classification.

(Prabavathy, Rathikarani, & Dhanalakshmi, 2020) concludes that the combination of MFCC and Sonogram features with SVM achieves the highest accuracy rate of 98% in classifying musical instruments. In contrast, Sonogram with kNN yields a lower accuracy of 95%. The proposed work demonstrates the superior performance of the SVM-based approach in this context. Future research directions may involve exploring deep learning techniques for musical instrument classification.

(Kour & Mehan, 2015) focuses on utilizing Mel-frequency Cepstral coefficients (MFCC) as a feature extraction method to classify musical genres based on short audio segments. The objective is to develop a music genre classification system employing Support Vector Machines (SVM) and Back Propagation Neural Network (BPNN) for effective categorization. The approach involves extracting acoustic features, including MFCCs, to characterize audio content and training SVM and BPNN classifiers using labeled data

the study also introduces an automated music genre classification system that leverages MFCC-based feature extraction and employs SVM and BPNN as classification algorithms. The primary objective is to classify audio content effectively into different genre categories.

(Kour & Mehan, 2015) objective of creating a music genre classification system by employing MFCC, SVM, and BPNN. From this study demonstrate the effectiveness of the proposed approach, which achieved a notable accuracy rate of 95% in music genre classification. Furthermore, the review highlights the comparative analysis with SVM, which exhibited an accuracy of 83%, emphasizing the superior performance of SVM and BPNN among machine learning algorithms in this context.

In this study, the primary objective is to explore the recognition of emotion conveyed through instrumental music, a relatively unexplored area within Music Information Retrieval (MIR). (Rajesh & Nalini, 2019), acknowledged the powerful role of music in evoking emotions and its widespread use in various applications, including music recommendation systems and music therapy. Recognizing the emotional content in music becomes increasingly important as the volume of digital music data grows.

The study focuses on the emotion recognition of instrumental music, specifically classifying emotions into four categories: happy, sad, neutral, and fear. To achieve this, the study employs a variety of acoustic features extracted from the music dataset. These features encompass Mel Frequency Cepstral Coefficients (MFCC), Chroma Energy Normalized Statistics (CENS), Chroma Short Time Fourier Transform (STFT), spectral features (spectral centroid, bandwidth, rolloff), and temporal features like Zero Crossing Rate (ZCR). (Rajesh & Nalini, 2019)

Deep learning techniques, particularly Recurrent Neural Networks (RNN), are used to train and recognize emotions based on these extracted features (Rajesh & Nalini, 2019). The study compares the performance of RNN with traditional machine learning algorithms, particularly Support Vector Machines (SVM), in the context of instrument emotion recognition.

The key findings of this study indicate that MFCC features combined with deep RNN provide the highest performance in instrument emotion recognition, achieving a recognition rate of 89.3%. Additionally, the study suggests that the type of musical instrument plays a role in determining the conveyed emotion within monophonic instrumental music clips.

(Rajesh & Nalini, 2019), review underscores the significance of recognizing emotions in instrumental music and presents a novel approach that leverages deep learning techniques for this purpose. The choice of acoustic features and the comparative analysis with traditional machine learning methods provide valuable insights into the effective recognition of emotions in instrumental music. Furthermore, the study hints at the potential for future exploration, particularly in the recognition of emotions in polyphonic instrumental music.

## Classification Using Deep Neural Networks

A common approach to applying CNNs to audio recognition tasks is employing a spectrogram image as the feeding data. However, this loses phase information. Park & Lee (2015) proposed using multiresolution recurrence plots (MRPs) to analyze time-series data in a two-dimensional space without losing phase information. The authors combine spectrogram images with MRPs using a multi-column network to improve classification performance over a system that uses only a spectrogram.

To evaluate the proposed method, the authors created a dataset classifying four different types of pianos using a single note. The dataset comprised four seconds of 88 single notes from each of the four different pianos without any audio compression or pitch shifting. The authors used ten-fold cross-validation to evaluate the performance of their proposed method.

The authors evaluated the performance of MRP-based classification, spectrogram-based classification, and the combined results using multi-column CNNs. The results showed that incorporating MRPs with spectrogram image data improved the classification performance. Using only network-1, the classification performance was higher than the baseline. However, spectrogram-based network-2 showed an even more improved performance than network-1. When combining MRPs with spectrogram image data using a multi-column network, the classification performance improved further.

In conclusion, the proposed method for musical instrument classification using CNNs and MRPs provides a significant improvement over traditional method. The combination of MRPs and spectrogram images using a multi-column network allows for the extraction of characteristic timbre of musical instruments that cannot be extracted using a phase-blinded representation such as a spectrogram. The proposed method shows promise for more challenging timbre classification for musical instruments (Park & Lee 2015).

In the field of music transcription, accurately deciphering the intricate details of polyphonic music signals has long been a challenge. However, a recent research paper titled "Multi-Instrument Music Transcription Based on Deep Spherical Clustering of Spectrograms and Pitchgrams" by Tanaka et al (2020). proposes a novel method that tackles this problem head-on. The paper introduces a clustering-based approach that estimates piano rolls of various musical instrument parts from complex polyphonic music signals, even when the musical instruments are undefined. The researchers' methodology consists of three key components: a feature extraction part, a feature embedding part for obtaining piano roll space and timbre space, and an estimation part based on deep spherical clustering. To address the limitation of dealing with undefined musical instruments, the team employs an instrument-independent neural multi-pitch estimator to estimate a condensed pitchgram. The pitchgram is then separated into distinct musical instrument parts using the deep spherical clustering technique. To improve transcription performance, the researchers propose a joint spectrogram and pitchgram clustering method that considers both timbral and pitch characteristics of musical instruments. The experimental results presented in the paper demonstrate the effectiveness of the proposed method. In the transcription of unknown instruments under open conditions, the proposed method outperforms the state-of-the-art classification-based approach. Moreover, the F-measure score for unknown instruments is comparable to that of known instruments, whereas the classification-based method experiences a significant decrease in performance. The proposed method also achieves accuracy equivalent to the classification-based approach in transcribing known instruments under both conditions (Tanaka et al 2020). The estimated piano rolls obtained using the proposed method exhibit promising results, effectively conducting pitch estimation and instrument assignment. Although some errors are present in certain cases, the overall performance of the method is commendable. The researchers highlight several areas for future exploration. One important direction is to automate the process of matching estimated piano rolls with instrument part labels, which is currently done manually. Additionally, the researchers suggest considering the possibility of multiple instruments sharing the same time-frequency bin by introducing the von Mises-Fisher (vMF) distribution into the hyper spherical latent space and employing soft clustering based on this distribution (Gopal & Yang, 2014). In conclusion, the research presented in this paper contributes significantly to the field of music transcription by providing a clustering-based method capable of transcribing arbitrary musical instrument parts. The integration of timbral and pitch characteristics enhances the accuracy and adaptability of the proposed approach. The results warrant further research, particularly in automating the matching process and exploring soft clustering techniques using the vMF distribution in the latent space (Tanaka et al 2020).

The recognition of predominant instruments in ensemble recordings poses a significant challenge, especially when distinguishing closely-related instruments such as alto and tenor saxophones. In their paper, "Jazz Solo Instrument Classification with Convolutional Neural Networks, Source Separation, and Transfer Learning, by Gómez, Abeßer, & Cano, (2018) from the Semantic Music Technologies Group at Fraunhofer IDMT in Germany propose an innovative approach to improve instrument recognition. Their study aims to make contributions to the field of Music Information Retrieval (MIR), benefiting tasks such as automatic music transcription, source separation, and music recommendation. The researchers build upon a hybrid deep neural network, which combines convolutional and fully connected layers to learn spectral-temporal patterns indicative of specific instruments.

To mitigate the overlap among multiple instruments, they evaluate two pre-processing steps harmonic/percussive and solo/accompaniment source separation algorithms (Gómez, Abeßer, & Cano, 2018). By isolating the desired instrument from the mixture, they aim to enhance instrument recognition performance. Additionally, they employ transfer learning techniques to fine-tune a pre-existing instrument recognition model for the classification of six jazz solo instruments. The results indicate that both source separation and transfer learning techniques significantly improve instrument recognition performance, particularly for smaller subsets of highly similar instruments. (Gómez, Abeßer, & Cano, 2018) found that the combination of solo/accompaniment source separation and transfer learning leads to better generalization to unseen data in jazz solo instrument classification. These findings demonstrate the potential of applying deep learning models to discriminate between highly similar instruments and extend their application to other timbre-related recognition tasks.

The systematic evaluation of source separation algorithms as pre-processing steps, along with the application of transfer learning techniques, contributes to the advancement of instrument recognition in ensemble recordings. The study provides insights into improving timbre description and instrument classification, particularly in jazz ensemble recordings. Furthermore, the findings suggest the potential utilization of the proposed system for content-based metadata clean-up and enrichment of jazz archives. However, (Gómez, Abeßer, & Cano, 2018) acknowledge the need for further research and emphasize the importance of expanding the dataset used in the study (JAZZ dataset). By increasing the dataset size, the generalizability of the system can be enhanced. Future investigations may explore the application of the proposed methods in performer identification and other timbre-related recognition tasks. In conclusion, Gómez, Abeßer, and Cano's study presents an effective approach to jazz solo instrument classification by combining convolutional neural networks, source separation, and transfer learning. Their research offers valuable insights and improvements to instrument recognition in ensemble recordings, paving the way for advancements in MIR and its related fields.

Music Instrument Classification (MIC) plays a vital role in various applications such as music recommendation, automatic mixing, and music discovery. However, the lack of annotated training data poses a significant challenge for achieving high-performance results. Chen and Lerch (2022), propose a novel technique called "reprogramming" to address the scarcity of annotated data for MIC. They demonstrate that reprogramming can effectively leverage the power of pre-trained deep neural networks originally designed for different tasks, resulting in state-of-the-art performance with significantly reduced training parameters.

Chen and Lerch (2022) adopt the concept of transfer learning and introduce the reprogramming paradigm to the field of Music Information Retrieval (MIR). They select a pre-trained state-of-the-art audio classification model, the Audio Spectrogram Transformer (AST), and extend it through input pre-processing and label mapping. The study aims to explore various forms of input and output reprogramming to enhance the compatibility of the pre-trained model with MIC tasks. The authors hypothesize that by modifying the input data and mapping the output labels, the reprogrammed model can adapt to the specific requirements of MIC.

Through extensive evaluation and analysis, the researchers demonstrate that reprogramming achieves remarkable performance in MIC tasks, surpassing state-of-the-art systems. The study provides evidence that reprogramming can be a promising technique for tasks impeded by data scarcity. By effectively utilizing a pre-trained model's representation, the reprogrammed system shows comparable or even superior performance while requiring significantly fewer training parameters. This approach addresses the limitation of data availability and reduces the complexity of training, making it applicable to various fields beyond MIR. In conclusion, the study by Chen and Lerch introduces the concept of reprogramming to address the lack of annotated data in Music Instrument Classification. Their methodology, which modifies the input and output of a pre-trained model, proves successful in achieving state-of-the-art results. The researchers highlight the potential of reprogramming in other MIR tasks and emphasize its low training complexity. Chen and Lerch (2022). They suggest exploring variations of input reprogramming beyond CNN or U-Net structures and testing the algorithm with different pre-trained models from both audio and non-audio domains. The promising outcomes and the reduced training complexity warrant further research into reprogramming as a transfer learning approach in MIR and other fields with limited data.

A proposed method for music emotion recognition based on a convolutional neural network (CNN). The method aims to improve the efficiency and accuracy of music emotion classification by combining low-level audio features with time-domain and frequency-domain features extracted using a convolutional recurrent neural network (CRNN) and bidirectional long short-term memory (Bi-LSTM) network. (Jia, 2022)

The proposed method utilizes the mel-frequency cepstral coefficient (MFCC) and residual phase (RP) as weighted and combined low-level audio features. These features are then input into the CRNN to extract time-domain, frequency-domain, and sequence features from the spectrogram. Additionally, the low-level features are input into the Bi-LSTM network to capture sequence information. The extracted features from both networks are fused and input into a softmax classification function with a center loss function to classify music into four emotion categories which are anger, happy, sad, and relaxation.

Experiments conducted on an emotion music dataset show that the proposed method achieves a recognition accuracy of 92.06% and a loss function value of approximately 0.98, outperforming other methods (Jia, 2022). The method offers a new approach for music emotion recognition by combining different types of features and leveraging deep learning techniques.

The passage also discusses the importance of music emotion recognition in various fields such as music database management, retrieval, recommendation, and therapy (Jia, 2022). It highlights the limitations of manual emotion labeling and the need for automatic recognition methods. Previous research in this area is mentioned, including studies on emotion recognition using lyrics or audio analysis, multimodal fusion techniques, and machine learning-based approaches.

The experiments conducted in the study use an emotion music dataset consisting of 2906 songs with four emotion classes: anger, happiness, relaxation, and sadness. The dataset is divided into training, verification, and test sets, with the first 30 seconds of each song used for analysis. The passage presents the results of different experiments, such as selecting the optimal audio time period, analyzing the influence of iteration times on loss values, comparing different convolutional networks, and evaluating emotion classification using confusion matrices.

Overall, the proposed method shows promising results in music emotion recognition and provides a new approach for extracting and combining various features for improved classification accuracy.

The classification of music genres plays a crucial role in effectively organizing and retrieving large collections of music. However, achieving reliable accuracy in music classification has remained a challenge. Traditional methods employing handcrafted features have struggled to capture the unique characteristics of music, necessitating the exploration of dynamic and effective alternatives. In this context, the combination of a Convolutional Neural Network (CNN) and variants of Recurrent Neural Networks (RNN) has been relatively underexplored. Ashraf et al., (2022) aims to address this gap by proposing a hybrid architecture that combines a CNN with Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), Gated Recurrent Unit (GRU), and Bidirectional GRU (Bi-GRU) variants of RNN. The performance of the proposed architecture is evaluated using two different feature extraction techniques: Mel-spectrogram and Mel-frequency cepstral coefficient (MFCC).

The dataset was preprocessed by converting the 30-second clips into 3-second durations for meticulous evaluation. Feature extraction was performed using Mel-spectrogram and MFCC methods. Mel-spectrogram represents a 2D signal obtained through bandpass filters and a digital filter bank, while MFCC involves analyzing audio signal acoustics such as rhythm, pitch, tonality, intensity, and timbre. The proposed hybrid architecture, incorporating CNN and the respective RNN variants, was then applied to classify the music genres.

The results showed that the hybrid architecture of CNN and Bi-GRU, utilizing Mel-spectrogram features, achieved the highest accuracy of 89.30%. On the other hand, the hybridization of CNN and LSTM, using MFCC features, attained an accuracy of 76.40% (Ashraf et al., 2022). These findings suggest that the proposed hybrid model is effective in music genre classification. The study contributes to the field by demonstrating the benefits of combining CNN and RNN variants, along with the importance of selecting appropriate feature extraction techniques. The authors also compared their model's performance with other state-of-the-art methods and found comparable results.

The hybrid architecture and feature extraction techniques presented in this study provide insights into improving music classification. However, further research is warranted to explore the proposed methodology on additional datasets, such as FMA, for tasks such as instrument recognition or artist identification. Additionally, it would be valuable to investigate the potential applications of the hybrid model beyond music classification, such as in music recommendation systems or online access platforms.

Overall, this study successfully addresses the challenges of music classification by proposing a hybrid CNN and RNN variant model. The experimental results, coupled with the comparative analysis, highlight the effectiveness of the proposed approach. By advancing the understanding of deep learning techniques in music analysis, this research contributes to the development of efficient systems for organizing and retrieving music collections.

In today's digital era, with the rise of music streaming platforms like Spotify, Apple Music, and Deezer, automatic instrument recognition in sound recordings has become increasingly important. However, distinguishing between similar instruments such as the cello and violin or the flute and clarinet remains a challenging task for both machines and humans. In this research paper titled "Instrument Classification using image-based Transfer Learning," Shukla et al., (2020). Aim to address this challenge by identifying four similar string instruments (acoustic guitar, cello, violin, and electric guitar) in audio recordings.

The researchers employed two popular deep learning architectures, 1D MLPs (Multi-Layer Perceptrons) and 2D CNNs (Convolutional Neural Networks), to classify the sounds. They compared the performance of these architectures on different audio features, namely Mel-frequency cepstral coefficients (MFCC), constant-Q transform (CQT), Mel Spectrogram, and a self-curated 1-dimensional feature that combines multiple audio features. Additionally, they explored the use of image-based transfer learning models such as Inception and VGG, which have shown significant advancements in image classification tasks.

The study's main contribution involved creating a subset of the IRMAS dataset, called IRMAS-String (Juan J. Bosch et al., 2012), specifically focusing on string instruments. The researchers conducted comprehensive experiments using various audio features and deep learning architectures for instrument classification. results indicated that utilizing pretrained image models, particularly Inception V3, led to improved accuracy compared to other architectures. They observed an increase in accuracy by converting stereo audio samples to mono using the Librosa package's to\_mono function. The study demonstrated the advantage of transfer learning in audio classification tasks and highlighted the effectiveness of MFCC and the self-curated 1D feature for instrument recognition. Shukla et al., (2020)

The research paper successfully achieved its objectives of comparing the performance of different audio features and model architectures in instrument classification tasks. The findings suggested that using image-based transfer learning models and specific audio features, such as MFCC, yielded promising results. The study's outcomes hold relevance in various domains, including music transcribers, home assistants, and recommendation systems. However, further research is warranted to explore the combination of different features and evaluate their collective performance. Shukla et al., (2020) Additionally, investigating preprocessing techniques like source separation and noise reduction could further enhance the efficiency of the models in instrument recognition.

Overall, the study sheds light on the potential of leveraging image-based transfer learning and deep learning architectures in the field of instrument classification. By addressing the challenges associated with distinguishing similar instruments, this research contributes to the development of accurate and efficient music classification systems.

In the Internet age, music has transcended borders and holds a universal appeal, expressing thoughts, emotions, and resonating with people from different cultures and regions. As technology revolutionizes various fields, including music production and education, researchers have started exploring the application of intelligent music recognition technology in music teaching. H., Zhang, Y., & Zhang, Q. (2022) investigated the use of deep learning, specifically the Long Short-Term Memory (LSTM) network, to distinguish and generate various genres of music. By analyzing the role of machine learning and deep learning in music, the study designs an algorithm model for intelligent music generation, providing a theoretical foundation for further research in this area.

The study utilizes a massive dataset of music to test the designed music style discrimination and generation model. The model's architecture consists of four hidden layers with 1,024, 512, 256, and 128 neurons, respectively. The classification accuracy of jazz, classical, rock, country, and disco music genres is evaluated, with a focus on jazz classification, which yields the best results at 77.5% accuracy. Additionally, the generated music scores are compared with the original music spectrum, demonstrating a close alignment. H., Zhang, Y., & Zhang, Q. (2022) employ deep neural networks, particularly LSTM, for music analysis and processing, as they are more capable of handling large-scale data and extracting meaningful features.

The experimental results indicate that the designed algorithm model can effectively distinguish between music signals and generate diverse genres of music. The classification accuracy for different music genres exceeds 60%, outperforming the traditional restricted Boltzmann machine method. The findings highlight the potential of deep learning algorithms, specifically LSTM networks, in music generation and classification tasks. The algorithm's ability to accurately capture and reproduce music characteristics paves the way for advancements in intelligent music teaching. The study's outcomes have implications for quality education and the integration of technology in music pedagogy.

This research contributes to the field of music education psychology by shedding light on the psychological changes that occur during music teaching. It emphasizes the importance of understanding individual behavioral and psychological cognition in the teaching process. The use of intelligent music signal identification and generation technology has the potential to enhance teaching effectiveness and provide a rich learning experience. Further research can focus on refining the algorithm model to improve accuracy, exploring additional music genres, and investigating the impact of intelligent music teaching on student learning outcomes. Continued research in this area will support the advancement of music education and the integration of technology in pedagogical practices.

The field of machine learning has witnessed a transformative shift with the emergence of Artificial Neural Networks (ANNs). This literature review explores recent developments in Convolutional Neural Networks (CNNs), a subset of ANNs specifically tailored for image-driven pattern recognition tasks. The intent of this review is to provide an overview of CNNs, highlight their significant features, and emphasize their relevance in simplifying complex image analysis tasks.

(O’Shea & Nash, 2015), begin by introducing ANNs as computational systems inspired by biological nervous systems. ANNs consist of interconnected neurons that collectively learn from input data to optimize their output (O’Shea & Nash, 2015). This foundational understanding sets the stage for exploring CNNs' unique characteristics.

CNNs are described as a subset of ANNs with a simplified yet powerful architecture. They are designed primarily for image recognition tasks. The review highlights the structure of CNNs, including input layers, hidden layers, and the process of learning through interconnected neurons. Deep learning, achieved by stacking multiple hidden layers, is also briefly mentioned (O’Shea & Nash, 2015).

The authors discuss two key learning paradigms in image processing: supervised and unsupervised learning. Supervised learning, involving labeled input data, is emphasized as the primary approach for image-focused pattern recognition tasks.

The review underscores the core similarity between CNNs and traditional ANNs in terms of self-optimization through learning. CNN neurons receive input data, perform operations, and generate a single perceptive score function. The last layer contains loss functions related to classes, and conventional techniques for ANNs remain applicable. The key distinction is that CNNs excel in image recognition due to their ability to encode image-specific features, reducing the model's required parameters (O’Shea & Nash, 2015).

The authors highlight a significant limitation of traditional ANNs, their struggle with the computational complexity of image data. They provide a comparison between the MNIST dataset (Deng, L., 2012), and larger colored images to illustrate the challenge. Larger images demand significantly more weights per neuron and a larger network, which can be impractical.

In conclusion, this literature review emphasizes the intent to introduce and clarify the role of CNNs in image analysis. It outlines the fundamental concepts of CNNs, discusses the necessary layers, and offers guidance on structuring CNNs for image analysis tasks. (O’Shea & Nash, 2015) aim to demystify the perception of complexity associated with these powerful machines learning algorithms, making them more accessible to beginners. Overall, this review underscores the significance of CNNs in the realm of image analysis and their potential to simplify complex problems by exploiting specific features of input data.

**HYPERPARAMETER TUNING**

This study delves into the crucial domain of hyper-parameter optimization (HPO) in machine learning (Yang & Shami, 2022). The study intent is to comprehensively explore the landscape of HPO, emphasizing its significance in enhancing the performance of machine learning models. The review covers essential features, including common optimization techniques, key hyper-parameters in ML models, state-of-the-art HPO methods, and practical applications. Additionally, it discusses existing libraries, frameworks, and emerging challenges in HPO research.

Machine learning has evolved as a dominant approach for solving data-related challenges across diverse domains (Yang & Shami, Year, 2022). To harness the full potential of machine learning models, optimizing their hyper-parameters has become imperative. This literature review aims to provide an extensive understanding of HPO, its methodologies, and its practical implications. The primary goal is to empower users, developers, analysts, and researchers in effectively tuning machine learning models by identifying the most suitable hyper-parameter configurations.

The study introduces the concept of HPO, emphasizing its critical role in enhancing machine learning model performance. It highlights the challenges associated with manual tuning and the need for automated optimization techniques.

(Yang & Shami, Year, 2022) elucidate the significance of hyper-parameters and distinguish them from model parameters. They discuss the types of hyper-parameters, including categorical, discrete, and continuous, and how they influence model behavior.

The study also explores various optimization techniques used for HPO, ranging from traditional methods like grid search and random search to more advanced approaches such as Bayesian optimization, multi-fidelity optimization, and metaheuristic algorithms. It discusses their strengths and applicability to different types of ML models.

The study presents an overview of available libraries and frameworks developed for HPO. These tools facilitate the practical implementation of HPO in machine learning projects, streamlining the optimization process.

(Yang & Shami, Year, 2022) share the results of experiments conducted on benchmark datasets, showcasing the comparative performance of different HPO methods. These practical examples illustrate the real-world impact of hyper-parameter optimization.

by discussing open challenges in HPO research and suggesting potential research directions. It encourages further advancements in HPO to address existing complexities and enhance ML applications.

This provides a comprehensive exploration of hyper-parameter optimization in machine learning. Its intent is to equip ML practitioners with the knowledge and tools to effectively tune models and improve their performance. By highlighting the features, challenges, and emerging trends in HPO research, this review serves as a valuable resource for the ML community, fostering better model development and understanding.

Alzheimer's Disease (AD), is a debilitating neurodegenerative condition that affects cognitive processes in older adults, ultimately leading to severe health decline and death. Early detection and accurate diagnosis of AD are critical for providing tailored medical treatment and interventions. This by (Veeralagan, J., & Manju Priya, S. 2022) explores the application of machine learning in the early diagnosis of AD, with a focus on hyper-parameter tuning to enhance the performance of machine learning models.

The study introduces the role of machine learning in healthcare, highlighting its potential to assist clinicians in diagnosing diseases. It underscores the importance of leveraging machine learning algorithms to analyze clinical data and improve diagnostic accuracy.

The study discusses the significance of hyper-parameter tuning in maximizing the performance of machine learning models. Various automated hyper-parameter tuning techniques are mentioned, including Randomsearch, Bayesian optimization, and Tree-structured Parzen estimators. However, the primary focus of this study is on GridSearchCV hyper-parameter tuning. (Veeralagan, J., & Manju Priya, S. 2022)

The review presents a range of machine learning algorithms suitable for early AD detection, including K-Nearest Neighbor (KNN), Gradient Boosting Classifier (GB), Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and Decision Tree (DT). Each algorithm is discussed in the context of their parameters and how GridSearchCV is used to optimize them

The study concludes that early detection of AD is crucial, and machine learning models with hyper-parameter tuning offer a promising approach. (Veeralagan, J., & Manju Priya, S. 2022), find that Support Vector Machine (SVM) with hyper-tuning is the most suitable model for early AD detection, achieving high accuracy, recall, and F1-score. While Decision Tree (DT) demonstrated high precision, other factors such as confusion matrix and ROC curve support SVM. The review suggests future research extensions involving other hyper-parameter tuning methods for further comparison and improvement in AD diagnosis.

(Bergstra, Yamins, and Cox 2013) primary focus is on the challenge of hyperparameter optimization in computer vision algorithms, which are integral for various tasks in the field. The authors emphasize that parameter tuning, often done manually, plays a critical role in achieving optimal algorithm performance. However, this process is not only time-consuming but also lacks formalization, making it difficult to assess whether an improvement in performance is due to better tuning or an inherently superior algorithm.

To address these issues, (Bergstra, Yamins, and Cox 2013), propose a meta-modeling approach aimed at automating hyperparameter optimization. The goal is to replace manual tuning with a systematic and reproducible optimization process

The results of this study showed that automated hyperparameter optimization can achieve state-of-the-art performance on diverse computer vision tasks. By systematically exploring the hyperparameter space, the approach outperforms manual tuning and provides reproducibility in model evaluation. They also emphasize the importance of efficient optimization algorithms and open-source software to support future research in hyperparameter optimization.

The study highlights the need for automated hyperparameter optimization in computer vision algorithms. It presents a structured approach that leverages optimization algorithms to achieve superior performance while addressing the challenges associated with manual tuning. The study's results demonstrate the effectiveness of this approach across various computer vision tasks and open up avenues for future research in this area.

(Chowdhury et al. 2022), focus revolves around the optimization of hyperparameters in the context of deep learning models. Hyperparameters play a crucial role in determining the performance of these models, and their optimal configuration is often a challenging task that requires in-depth knowledge of deep learning algorithms and hyperparameter optimization (HPO) techniques. While various automatic optimization approaches exist, each has its own set of advantages and limitations, making it essential to understand their performance across different datasets and architectures. (Chowdhury et al. 2022)

The study addresses the need for efficient hyperparameter optimization in deep learning, emphasizing the importance of fine-tuning these parameters to match specific datasets and achieve optimal model performance.

To assess the performance of HPO algorithms, the study considers a range of commonly used optimization techniques, including Grid search (GS), Genetic algorithm (GA), Bayesian optimization (BO), Random search (RS), Hyperband (HB), and Particle swarm optimization (PSO).

(Chowdhury et al. 2022) aims to determine whether the performance of HPO algorithms remains consistent when applied to different datasets and architectural configurations, shedding light on their reliability and generalizability.

The motivation for this study arises from the growing use of deep neural networks in real-world applications, where large datasets make manual hyperparameter tuning impractical. The focus is on automating this process to enhance the efficiency of deep learning models.The study highlights the importance of using consistent hyperparameter optimization methods when comparing different deep learning algorithms, ensuring fair and reproducible evaluations.

# VALIDITY

Validity, in the context of research, refers to the extent to which a study or experiment accurately measures or addresses what it intends to. It encompasses both accuracy and relevance. In other words, a valid study or model is one that accurately and appropriately assesses and addresses the research problem or question. The components of validity chosen for this study are

Accurate: This component is relevant as the deep learning model's output needs to be statistically sound and secure. The accuracy of the model's output is critical to the success of the classification and analysis of musical instruments. It also refers to how well the deep learning model's output matches the actual, ground truth data. In the context of classifying musical instruments, accuracy is crucial because it measures the model's ability to correctly identify and classify different instruments based on the provided data. High accuracy indicates that the model is making fewer errors, which is essential for the success of any classification task.

Relevant: This component is important because the information collected through spectrograms and audio features should directly relate to the problem definition identified, which is the accurate classification of musical instruments. The relevance of the features used in the model will determine its effectiveness in classifying different musical instruments. Relevance focuses on the appropriateness of the features and data used in the model. In this case, it ensures that the spectrograms and audio features employed directly contribute to the problem at hand, which is classifying musical instruments. Irrelevant features could introduce noise and hinder the model's performance. Ensuring that the features used in the model are relevant is essential to avoid overfitting and to optimize the model's performance. Features that are unrelated to the problem could lead to a less effective model.

# Sampling Strategy

For a data analytics project focused on classifying and analyzing musical instruments, specifically snowballing and judgment sampling, has been chosen. This approach offers several advantages, such as flexibility in data collection, targeting specific elements of the population, and leveraging expert knowledge. Snowballing involves gradually expanding the sample by asking participants to refer others who meet the desired criteria, while judgment sampling allows for selection based on expert judgment and domain knowledge (Parker, Scott, & Geddes, 2019). These methods align with the project's objectives of developing a deep learning model for instrument classification and ensure a diverse and comprehensive dataset. Overall, this sampling strategy enhances the accuracy and effectiveness of the deep learning model and contributes to music classification and analysis.

# Primary Research Methodology

A qualitative research method, specifically in-depth interviews, have been chosen as the primary research methodology. These interviews provide rich qualitative data, capturing detailed aspects of instrument classification that may not be captured by quantitative methods alone. The interviews revealed potential challenges and limitations in instrument classification, such as cases where instruments produce similar sounds or variations in musical interpretation. This awareness guided the research in addressing these challenges effectively. The flexibility of in-depth interviews enables tailored questioning and exploration of specific areas of interest, accommodating the diversity within the field of music. Furthermore, these interviews foster a personal connection, encouraging participants to share their authentic perspectives and insights. Overall, the choice of in-depth interviews as the primary research methodology will enhance the accuracy and reliability of the deep learning model and contribute valuable insights to the field of music information retrieval and analysis.

# Ethics

Ethical considerations are essential in a data analytics project on classifying and analyzing musical instruments. Key considerations include privacy and data protection, fairness and bias mitigation, transparency and explain ability, impact on individuals, and responsible dissemination of findings. Compliance with data protection regulations, such as the General Data Protection Regulation (GDPR) in the European Union, is essential. These regulations outline specific requirements for the collection, storage, and processing of personal data. Researchers must ensure that they adhere to these regulations when handling any personal data. ("Article 6: How to Process Personal Data Legally," GDPR.eu)

Protecting the privacy and confidentiality of participants is a primary concern. Anonymizing interview data is a common practice to ensure that participants cannot be personally identified. Removing or replacing any personally identifiable information is essential before analysis. ("Article 6: How to Process Personal Data Legally," GDPR.eu)

It is crucial to obtain informed consent, handle data securely, and communicate the purpose and scope of data collection to participants

Prior to conducting interviews or collecting any data, obtaining informed consent from participants is crucial. Participants must be fully aware of the project's purpose, how their data will be used, and any potential risks or implications. This practice aligns with GDPR standards.

# METHODOLOGY

The methodology section of this project outlines the approach and techniques employed to accomplish the task of classifying and analyzing musical instruments based on spectrograms and audio features. The primary objective is to design a robust deep learning model that can accurately identify and differentiate between various musical instruments, thereby contributing to the advancement of music information retrieval and classification systems. This section provides a comprehensive overview of the steps taken to achieve this goal, from data preprocessing and feature extraction to model architecture and evaluation metrics.

The methodology encompasses a multi-faceted approach, integrating state-of-the-art techniques in deep learning, signal processing, and machine learning. The design decisions are guided by the need to address challenges such as spectral similarities between instruments, data variability, and the extraction of discriminative features from audio recordings. By combining these techniques, we aim to achieve a high-performing model capable of classifying musical instruments with a high degree of accuracy.

The following subsections elaborate on the key components of the methodology, including data preparation, feature extraction, deep learning model architecture, training and evaluation, and the rationale behind the chosen approaches. Through a detailed exploration of each phase, this methodology provides insights into the workflow and considerations that underpin the development of an effective musical instrument classification system.

# DATASET

The foundation of this project lies in the utilization of the IRMAS dataset (Juan J. Bosch et al., 2012). Which is under the Creative Commons Attribution 4.0 International License (CC BY 4.0), the dataset has a collection of musical audio excerpts meticulously annotated with the primary instrument(s) present in each recording. This dataset serves as the cornerstone for evaluating the efficacy of our proposed methodology. The dataset was previously employed for assessment in a study titled "A Comparison of Sound Segregation Techniques for Predominant Instrument Recognition in Musical Audio Signals" by Bosch et al. (2012), showcased in the proceedings of the International Society for Music Information Retrieval (ISMIR).

IRMAS stands as a pivotal resource for training and validating methods aimed at the automatic recognition of predominant instruments within musical audio. The diverse set of instruments encompassed by IRMAS includes cello, clarinet, flute, acoustic guitar, electric guitar, organ, piano, saxophone, trumpet, violin, and human singing voice. Notably, this dataset diverges from its precursor, compiled by Ferdinand Fuhrmann, in a few key aspects. The dataset undergoes meticulous data preparation, including stereo audio conversion, to ensure compatibility with the project's objectives. The subsequent sections delve into the methodology's intricate components, encompassing feature extraction, model architecture, training, and evaluation

# FEATURE EXTRACTION

The heart of our feature extraction methodology lies in the utilization of spectrograms, providing a profound visual representation of audio signals frequency content over time. By deconstructing audio into its constituent frequencies, spectrograms unveil the dynamic evolution of these frequencies across time. This transformative representation adeptly captures variances in timbre, harmonics, and other pivotal sonic attributes. (Bartusiak & Delp, 2022)

For the purpose of our study, the generation of spectrograms is paramount. Each audio excerpt within our dataset is processed to create spectrograms. These spectrograms function as intricate maps that detail how the frequency components unfold throughout the duration of the audio. Consequently, they offer a window into the tonal nuances that distinguish one instrument from another. These spectrograms establish a bedrock for our ensuing feature extraction endeavors. (iZotope, 2020)

In tandem with spectrograms, we delve into the extraction of Mel-frequency cepstral coefficients (MFCCs) from the raw audio data. This process involves a series of steps that transforms the continuous audio signal into numerical representations. These numerical MFCC values hold a wealth of information about the spectral characteristics of the audio. Racharla et al., (2020).

The process of extracting MFCCs begins with segmenting the audio into short time frames. Each frame is then subjected to a Fourier transform to convert it from the time domain to the frequency domain (Yang et al., 2020). Subsequently, the power spectrum is mapped to the Mel scale, which is closely aligned with human auditory perception. This Mel-frequency representation is then subjected to a discrete cosine transform (DCT), yielding the final MFCCs. Each MFCC coefficient encapsulates a different aspect of the audio's spectral content, contributing to a holistic portrayal of its timbral intricacies. Zahid et al., (2015)

As an integral step, the extracted MFCC coefficients are organized into a structured data frame. This data frame serves as a structured repository that houses the MFCC values for each audio excerpt, effectively creating a digital representation of the audio dataset. Each row corresponds to an audio clip, with the columns containing the MFCC coefficients for that clip. This structured format enables streamlined data management and facilitates seamless integration with machine learning algorithms.

The journey of feature extraction doesn't halt with MFCCs. Building upon this foundational audio representation, The MFCCs are transformed back into spectrograms. By aligning numerical MFCC data with its spectrogram counterpart, the gap between the mathematical and visual domains, offering a comprehensive perspective on audio characteristics.

In this multifaceted approach, the numerical MFCC values not only serve as crucial inputs for machine learning but also empower the visualization of audio attributes. The conversion from MFCCs to spectrograms completes the feedback loop, enriching and understanding of the intricate connections between mathematical representations and sonic qualities.

In summary, the feature extraction process combines the visually informative power of spectrograms with the numerical precision of MFCCs. While spectrograms offer a panoramic view of frequency changes over time, MFCCs distill complex audio into concise numerical representations. These numerical values, carefully organized into data frames, enable seamless integration with machine learning models. Through the extraction of MFCCs, their conversion into spectrograms, and their organization in structured data frames, the methodology unlocks a holistic understanding of the diverse sonic landscapes inhabited by different musical instruments.

# Model Architecture

To classify musical instruments using deep learning led to design a robust and accurate model architecture. Leveraging both spectrograms and extracted audio features, the architecture utilizes the power of Convolutional Neural Networks (CNNs) to process spectrograms, coupled with other models for handling audio features. In this section, elucidates the high-level blueprint of the model, the exploration of alternative techniques, and the journey toward optimizing accuracy. The primary deep learning architecture centers around a CNN tailored for processing spectrograms. Spectrograms, capturing the temporal and spectral nuances of audio, serve as rich data inputs for the model. The architecture comprises layers for convolution, pooling, and fully connected (dense) layers. Convolutional layers convolve over spectrogram data, effectively learning high-level features. Pooling layers aid in reducing the spatial dimensions while retaining important information. (Rodriguez, 2022)

Each convolutional layer is followed by an activation function, often Rectified Linear Units (ReLU), which introduces non-linearity to the model. To prevent overfitting, a Dropout layer is strategically inserted, randomly deactivating neurons during training (Rodriguez, 2022). Batch normalization ensures stable training dynamics by normalizing activations, and MaxPooling layers capture essential information while reducing computational complexity. (Dalyac, Shanahan, & Kelly, 2014). The final layers consist of fully connected (dense) layers, culminating in the output layer with neurons corresponding to the number of instrument classes. A SoftMax activation function generates class probabilities, facilitating accurate predictions.

While the primary architecture exploits spectrograms, the exploration extended to incorporate models utilizing the extracted MFCC features. Support Vector Machine (SVM), Decision Tree, and Random Forest models were considered for comparison (Racharla et al., 2020). SVM, a powerful classification technique, exploits hyperplane separation to categorize data points. It offers variations like linear, polynomial, and radial basis function (RBF) kernels. SVM, a versatile classification algorithm, was deployed using the MFCC features that was extracted. (Racharla et al., 2020). The linear kernel, capable of separating linearly separable data, and the RBF kernel, adept at capturing complex relationships, were particularly promising. In the pursuit of accuracy, A multifaceted evaluation approach was embraced. The model's success isn't solely gauged by overall accuracy, as it may not fully represent class imbalances or false positives. Instead, examined precision, recall, F1 score, and the confusion matrix. Precision captures the ratio of true positive predictions to all positive predictions, offering insights into false positives. Recall denotes the ratio of true positives to all actual positives, illuminating false negatives. The F1 score harmonizes precision and recall, providing a comprehensive metric (Racharla et al., 2020).

In the pursuit of constructing a comprehensive instrument classification system, the realm of decision trees and their more robust counterpart, random forests, was delved into. These techniques offered both interpretability and the power of ensemble learning, rendering them fitting candidates for the intricate task of music instrument recognition.

A decision tree is a hierarchical structure that makes decisions based on feature values (Safavian & Landgrebe, 1991). It's analogous to a flowchart, where each internal node represents a feature, each branch signifies a decision, and each leaf node represents an outcome. Decision trees excel at handling non-linear relationships and are adept at classifying multi-class data, making them well-suited for the diverse range of musical instruments. (Afram & Sarab Fard Sabet, 2023).

In the context of music instrument classification, decision trees break down the complex decision-making process into a sequence of straightforward choices. Each split considers a particular feature to segregate data into subsets. This hierarchical approach aligns with the distinct characteristics of musical instruments. For instance, the decision tree might identify a crucial feature like the spectral envelope or harmonic content to distinguish between a flute and a guitar.

Building upon the foundation of decision trees, random forests elevate accuracy and robustness through ensemble learning. A random forest comprises multiple decision trees, each trained on a different subset of the data and features (IBM, 2023). The outputs of individual trees are aggregated to yield a more accurate and stable prediction.

In instrument classification task, the random forest's ensemble nature mitigates the risk of overfitting and enhances generalization to unseen data (Racharla et al., 2020). By aggregating the decisions of multiple trees, random forests reduce the impact of noisy or inconsistent data points.

One compelling aspect of decision trees and random forests is their inherent interpretability. Decision trees offer a clear insight into how a model arrives at a decision. By tracing the path from the root node to the leaf node, the sequence of features and their thresholds that led to a classification can be discerned.

Additionally, random forests provide a quantification of feature importance. This assessment highlights which features played the most significant role in distinguishing instruments. Understanding feature importance is not only valuable for interpreting the model's decisions but also aids in refining future feature extraction and model design.

Ventured into using decision trees, random forests and SVM, the focus extended beyond mere accuracy. The imbalance in the dataset, with varying numbers of audio samples for different instruments, necessitated a nuanced evaluation. To address this, confusion matrix was employed (Racharla et al., 2020), which provided a detailed breakdown of true positive, true negative, false positive, and false negative predictions.

In addition to the confusion matrix, the F1 score and precision as key evaluation metrics were also adopted. The F1 score harmonizes precision (minimizing false positives) and recall (minimizing false negatives), offering a balanced assessment. Precision, on the other hand, highlights the ratio of true positive predictions to all positive predictions, effectively minimizing false positives. These metrics ensure that the model's accuracy is not biased by class imbalances and false predictions, (Racharla et al., 2020).

While the extracted Mel-frequency cepstral coefficients (MFCC) provide valuable insights into the audio features of the musical instruments, the exploration extended by delving into the realm of spectrograms.

Spectrograms break down audio signals into a grid of time versus frequency, with color intensity indicating the amplitude of different frequency components (Shukla et al 2020). This visual representation allows us to witness the dynamic interplay of various frequencies across time, providing an invaluable tool for analyzing tonal characteristics, harmonics, and transient elements within the music. (Shukla et al 2020)

Unlike MFCC, which compresses audio information into a limited set of coefficients, spectrograms present a comprehensive snapshot of the frequency content. This enables, to capture both subtle variations and sudden shifts in musical timbre and texture. Spectrograms serve as an ideal foundation for neural network models, as they encapsulate the essence of the audio signal and empower models to discern intricate patterns and relationships.

With an understanding of spectrograms, their potential is harnessed in constructing robust neural network models for music instrument classification. Embarking on building two distinct architectures: Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN). These architectures tap into the essence of spectrograms to decipher the complex audio information encoded within.

ANNs are versatile models designed to mimic the human brain's decision-making process (Haykin, 2009). ANNs consist of interconnected layers, including input, hidden, and output layers. Each neuron processes input and passes it to subsequent layers through weighted connections. ANNs are adept at recognizing patterns but might lack the spatial awareness to discern finer spectral nuances.

CNNs, on the other hand, exhibit remarkable prowess in spatial recognition, making them ideally suited for image-like data such as spectrograms. Inspired by the human visual system, CNNs employ convolutional layers to extract features from local regions of the input. These features capture spatial hierarchies, enabling CNNs to identify complex patterns that ANNs might overlook.

In the pursuit of constructing ANN and CNN models, partition our dataset into training, validation, and testing subsets. This division ensures that the models are not only trained on a diverse range of data but also tested on unseen samples. Proceeding with model training, the confusion matrix will be the guiding light, providing insights into the distribution of correct and incorrect predictions across instrument classes.

By transitioning from MFCC to spectrograms and embracing neural network models like ANN and CNN, aim to unearth the profound complexities embedded within audio signals. Spectrograms provide a visual symphony of audio dynamics, enriching the feature representation (Agrawal, 2023). ANN and CNN architectures capitalize on this representation, leveraging their unique strengths to decode the intricate world of musical instruments. Venturing into the realm of neural networks, the journey is underscored by the pursuit of accuracy, interpretability, and an unwavering commitment to unveiling the resonant essence of music.

Top of Form

Bottom of Form

Incorporating decision trees and random forests into the classification pipeline brought interpretability, robustness, and meticulous evaluation to the music instrument recognition endeavor. The hierarchical structure of decision trees aligned with the intricate distinctions among musical instruments, while the ensemble nature of random forests bolstered accuracy and generalization. Through these methodologies, embarking on a journey to construct a versatile model that not only delivers accurate predictions but also aids in unraveling the intricate tapestry of musical timbre.

Top of Form

Bottom of Form

# IMPLEMENTATION

In the implementation, several key decisions were made to streamline the preprocessing of the musical instrument dataset. Among these, the careful renaming of audio files and the thorough format verification were crucial steps that contributed to the efficiency and effectiveness of the project. Let's delve into why these decisions were made and how they impacted the overall implementation.

In the initial stages of handling a dataset as diverse as the IRMAS collection, organization is paramount. Renaming the audio files served multiple purposes. The original filenames might be diverse, inconsistent, and not directly indicative of the content they represent. Renaming them according to a standardized format, using instrument labels provided a consistent naming convention across the dataset. With a clear and informative naming structure, locating and referencing specific audio files during implementation, training, and evaluation became much more straightforward.

Renaming files programmatically according to instrument labels saved manual effort and mitigated the risk of human errors. Organizing files by instrument labels enhanced the dataset's integrity, allowing to effortlessly manage and track the data's evolution.

Working with a well-defined and uniform file format is essential for consistent processing. The choice of using WAV (Waveform Audio File Format) files was driven by several considerations.WAV files are known for their lossless audio quality. This fidelity was crucial for preserving the nuances and subtleties of musical instruments during preprocessing and analysis. Many audio processing libraries and frameworks readily support WAV files (Whibley, S. 2016). This compatibility simplified the extraction of features, generation of spectrograms, and model training.

In the implementation process, one of the initial steps involved checking the sample rate of the audio files in the dataset. While the sample rate was not actively converted, the fact that the audio files uniformly exhibited a sample rate of 22.05 kHz holds significant implications for the compatibility and quality of our subsequent analyses. Let's explore why verifying and maintaining this sample rate is advantageous for the project. A uniform sample rate simplifies the preprocessing pipeline. Audio processing libraries and tools commonly expect a consistent sample rate, enabling seamless integration without the need for extensive conversions. When the sample rate is consistent, it ensures that all audio excerpts are compatible with the same processing steps, also avoiding the complexities that arise from dealing with multiple sample rates. the fact that the dataset already adheres to a consistent sample rate of 22.05 kHz is a noteworthy advantage. It ensures compatibility, preserves audio quality, and sets the stage for accurate and insightful analyses of musical instrument characteristics. This consistency underpins the reliability and effectiveness of the implementation process.

Transitioned to working with the extracted Mel-frequency cepstral coefficients (MFCC), thoughtfully organized into a structured data frame. This data frame featured 21 columns: the first column housing the instrument names and the subsequent 20 columns containing the stored MFCC values. With these foundational elements in place, the exploration embarked on a multifaceted journey, utilizing these MFCC values as the cornerstone for building and evaluating a range of models, including Support Vector Machines (SVMs), Decision Trees, and Random Forests.

## SVM DECISION TREES AND RANDOM FORESTS

A strategic move to adopted a technique of randomly selecting 400 MFCC coefficients from each class label. By doing so, introduced an element of variability that captured different aspects of the audio's spectral content.

For achieving robust musical instrument classification, embarked on a journey starting with the implementation of a Linear Support Vector Machine (SVM) model. This model leverages extracted MFCC values as input features, paving the way for insightful insights and enhanced accuracy. Let's delve into the steps and discoveries made during this phase of exploration. (Racharla et al., 2020)

The journey commenced with extracting instrument labels from the structured data frame. These labels, initially in string format, needed to be transformed into numerical values for compatibility with the SVM model (Racharla et al., 2020). To achieve this, we introduced a crucial ally the Label Encoder (Rodríguez et al., 2018). This component deftly converted the instrument labels into numerical representations, facilitating seamless integration into the SVM model.

The next pivotal step involved transforming MFCC values into a NumPy array of floating-point numbers, a foundational requirement for the SVM model. But this wasn't a simple transformation; we introduced the Standard Scaler (Ahsan et al., 2021). This element of the pipeline standardized the MFCC data by centering it around zero and scaling it to have unit variance. Why is this crucial? Scaling eliminates variations in magnitude across features, ensuring that no particular feature dominates the others. This process equips the SVM model to treat all features with equal importance, enabling more effective classification.

With the standardized MFCC data in place, the next move involved splitting the data into training and testing subsets. This division facilitates thorough model evaluation and ensures that the model's performance isn't merely memorizing the training data. 20% of the data is allocated for testing purposes, while the remaining 80% became the training set, primed to teach the model the intricacies of instrumental sound. (Racharla et al. 2020, Chen and Lerch 2022).

The heart of the Linear SVM model lies in its kernel. A kernel acts as a transformation that maps the original input data into a higher-dimensional space, enhancing the model's ability to capture complex relationships. For this initial implementation, the linear kernel, renowned for its simplicity and efficacy in handling linearly separable data, was opted for.

A critical parameter in SVM models is the regularization parameter, often denoted as C. This parameter governs the trade-off between maximizing the margin between classes and minimizing the classification error. A lower value of C encourages a wider margin but allows some misclassifications, while a higher value of C enforces a narrower margin to minimize misclassifications (Yang & Shami, 2022). Additionally, to ensure the best possible performance, we employed a grid search cross-validation (GridSearchCV) approach, which allowed us to systematically explore various C values and other hyperparameters to find the optimal configuration for our SVM models (Chowdhury et al. 2022)

With the Linear SVM model crafted, trained, and tested, we turned our gaze to the results. The accuracy of any model serves as a crucial benchmark for its performance. In this initial iteration, the Linear SVM achieved an accuracy of 62%, signifying a promising start. Through the rigorous grid search cross-validation (GridSearchCV) process, determined that the best parameters for the model were {'C': 1, 'kernel': 'linear'}.

Moving forward, the journey expanded to explore SVMs with different kernels, eager to optimize our model's accuracy. We introduced polynomial and radial basis function (RBF) kernels into the equation. The polynomial kernel, capable of capturing non-linear relationships, elevated the accuracy to 76%. However, it was the RBF kernel that outshined the rest, delivering an accuracy of 81%.

In summary, the foray into SVMs proved enlightening and fruitful. The Linear SVM provided a solid foundation, offering valuable insights into the data's linear separability. The subsequent introduction of polynomial and RBF kernels illuminated the potential of capturing complex relationships among instrument features (Racharla et al., 2020). The accuracy progression across these models highlighted the role of kernel selection in model performance.

The Linear SVM, coupled with its polynomial and RBF counterparts, sets the stage for further exploration into advanced neural network models, where the synergy of spectrograms and MFCC values promises to unlock deeper layers of audio patterns and characteristics. the pursuit of excellence continues as the boundaries of musical instrument classification are pushed and, armed with newfound insights and a thirst for innovation.

The exploration for the optimal instrument classification model continued, Setting sights on the realm of ensemble learning, specifically the Random Forest classifier. This technique weaves together multiple decision trees to create a powerful ensemble model, offering a harmonious blend of accuracy and robustness. Let's delve into the implementation of the Random Forest classifier, the configuration choices made, and the resonating results achieved.

The cornerstone of this endeavor was the Random Forest classifier, an ensemble learning technique that unites the strengths of multiple decision trees. Each decision tree contributes its own insights, and their collective wisdom enhances the model's accuracy and generalization capabilities. The essence lies in the diversity and synergy of these individual trees, which together compose a cohesive and potent classification system.

The importation of the Random Forest classifier, setting the stage for its configuration. Among the key parameters, the estimator count significantly influences the model's performance. For the implementation, opted for 150 decision trees after applying a grid search cross-validation (GridSearchCV), (Chowdhury et al. 2022). This choice strikes a balance between model complexity and computational efficiency, ensuring accurate predictions without overwhelming the system.

Reproducibility and consistency in machine learning experiments are paramount. To ensure the results could be replicated, A random state of 7000 was introduced. This seed value serves as a guiding light, orchestrating the randomness within the model in a consistent manner. By anchoring the randomness, A controlled environment is created for experimentation and evaluation.

With the configuration in place, model training was embarked upon., allowing the Random Forest classifier to learn from the data and distill patterns that distinguish between musical instruments. As the training progressed, the model drew insights from the diverse features and nuances encoded in the MFCC data.

The culmination of this process was an accuracy of 0.81. This accuracy level attests to the model's capability in identifying the intricate characteristics that set different instruments apart. The resonance of this achievement reflects the powerful synergy of ensemble learning and the holistic understanding of instrument nuances encoded within the MFCC values.

Embarking on the final leg of the journey through classification techniques, attention was turned to the venerable Decision Tree model. This model, built upon the fundamental principles of partitioning and hierarchy, offers a glimpse into how decisions are crafted from data. the implementation of the Decision Tree model delved into the nuances of the Gini Index, a parameter that resonates with its unique learning approach. Let's explore the essence of this parameter, the model's decision-making process, and the symphony of accuracy that ensued.

Central to our Decision Tree model, with the help of grid search cross-validation was the Gini Index, a parameter that encapsulates the concept of impurity. In essence, the Gini Index measures the likelihood of misclassification by evaluating the diversity of classes within a node. A lower Gini Index value indicates a purer node, where all instances belong to a single class, while a higher value signifies greater impurity due to the mixture of classes. the choice of the Gini Index as the criterion for decision-making is rooted in its effectiveness in crafting decision trees that achieve optimal separation of classes.

At the core of the Decision Tree's journey lies its ability to make decisions through a hierarchical structure. During the training process, the model learns to partition the data based on the provided features. The process entails evaluating the Gini Index at each potential split, seeking to minimize impurity and maximize class separation. As the model traverses the data, it constructs a tree of decisions, where each internal node represents a feature, each branch signifies a decision, and each leaf node embodies an outcome.

As the Decision Tree model completed its training, it emerged as a testament to the power of hierarchical decision-making. Its accuracy of 71% echoed with the symphonies of correctly classified instances, harmonizing with the nuances of the musical instrument dataset. The Gini Index, with its focus on impurity reduction, guided the model to craft decisions that effectively separated instrument classes.

With Decision Trees, Random Forests, Support Vector Machines, and ensemble learning within the repertoire, standing poised on the cusp of delving into the realm of advanced neural networks, the journey, marked by curiosity and a passion for musical analysis, marches forward in seeking to uncover the hidden melodies encoded within the heart of data.

## NEURAL NETWORKS

In this phase, delving into the intricate domain of neural networks, with a specific focus on Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN). These networks stand as sophisticated computational architectures that exhibit the capacity to comprehend and categorize musical instruments based on visual representations of their acoustic profiles, known as spectrograms (AltexSoft & Brux lab. 2022).

The initiation into the neural network landscape begins with the meticulous curation of spectrograms, which serve as graphical representations of audio content. In ensuring a harmonious input structure, all spectrograms are meticulously resized to a standardized dimension of 244x244 pixels (O’Shea & Nash, 2015). This uniformity not only fosters visual consistency but also streamlines processing efficiency within the network architecture

From the abundance of spectrograms, a judicious curation ensued, resulting in the selection of 400 images for each musical instrument. This deliberative assembly constitutes the foundation for training the neural networks. In this process, two distinct repositories were initiated, designed to house both the image data and their corresponding labels.

The neural network's takes shape in the form of a multi-dimensional array, comprising 2000 samples, each with dimensions of 244x244 pixels and a color spectrum represented by three channels. This composite array, measuring 2000x244x244x3, serves as the primary input for the neural networks. Within these arrays lies a visual representation of musical instruments, providing a diverse array of inputs to engage the networks' classification capabilities.

Much like translating the notes of a musical composition into a universal language, the labels assigned to various musical instruments undergo a process known as one-hot encoding (Rodríguez et al., 2018). This transformation empowers neural networks to discern the categorical essence embedded in these labels. Each instrument label evolves into a numerical composition, seamlessly harmonizing with the architecture of the network and facilitating nuanced differentiation across the musical instrument spectrum.

The curated dataset is partitioned into three distinct subsets to facilitate the training, validation, and testing of the neural networks. The training subset, encompassing 1280 spectrogram images, serves as the foundational phase wherein the networks internalize patterns. The validation subset, comprising 320 images, functions as a controlled environment for fine-tuning and optimization. Finally, the testing subset, comprising 400 images, offers an unbiased evaluation of the networks' acquired proficiency.

In the realm of Convolutional Neural Networks (CNNs), Embarking on crafting an architecture proficiently deciphering the intricate spectrogram representations of various musical instruments. The CNN journey unfolds through a series of strategic layers, each serving a pivotal role in the classification process.

The CNN's blueprint comprises three convolutional layers, each fortified with Rectified Linear Unit (ReLU) activation (Sakib et al., 2019). These layers act as perceptive filters, discerning distinct features within the spectrogram images. Accompanying them are max pooling layers (Sakib et al.,2019, O’Shea & Nash, 2015), which play a crucial role in dimensionality reduction. These layers spotlight salient features while dismissing redundant information, a pivotal step in feature extraction.

The narrative evolves with the introduction of a flattening layer. This layer is instrumental in reshaping the extracted 2D features from the convolutional and pooling layers into a streamlined 1D vector (Jeczmionek & Kowalski, 2021). This transition sets the stage for the subsequent layers' integration, a unifying bridge between the convolutional insights and the forthcoming dense layers.

A dense layer with 64 units assumes its place in the architecture (Jeczmionek & Kowalski, 2021), cultivating a network of interconnected neurons. This layer's role is to distill the amalgamated features into a more concise yet comprehensive representation. Introducing a dropout layer, set at 0.5, bolsters the network's resilience against overfitting. Dropout selectively deactivates neurons during training, effectively curbing the network's reliance on specific paths, thereby enhancing generalization. (Srivastava et al., 2014)

The final dense layer, adorned with 5 units, corresponds to the instrument classes we seek to identify. Employing a SoftMax activation function (Sakib et al.,2019), this layer generates probabilities for each class, enabling precise classification in multiclass scenarios. In tandem, we harnessed a learning rate of 0.001 and an Adam optimizer, a dynamic optimization algorithm renowned for efficient convergence and adaptability. (Jeczmionek & Kowalski, 2021)

The choice of categorical cross entropy as the loss function is predicated on its compatibility with multiclass classification tasks. This loss function gauges the disparity between predicted and actual class probabilities, aligning seamlessly with our objective of instrument classification across multiple classes. (Jeczmionek & Kowalski, 2021)

Incorporating early stopping augments our CNN's training process. This technique monitors the validation loss and halts training once the loss ceases to decrease, preventing overfitting to the training data. The introduction of early stopping underscores the commitment to fostering a network that generalizes beyond the training set. (Prechelt, 2002)

Notably, manual hyperparameter tuning, often referred to as "babysitting," (Yang & Shami, 2022, p. 23) played a significant role in the model's success. Due to issues with Keras Tuner, a hyperparameter tuning library, manual adjustments were meticulously made to fine-tune the network's parameters. This painstaking process involved adjusting hyperparameters such as the number of units in layers and learning rates, aiming to optimize model performance.

the CNN is nurtured through a training process characterized by a batch size of 32 and 10 epochs. This orchestrated training sequence converges to an astounding accuracy of 99%, a testament to the efficacy of the architectural design and training strategy.

The narrative culminates in a meticulous assessment using the confusion matrix. This matrix dissects predictions, revealing true positives, true negatives, false positives, and false negatives. This nuanced evaluation strategy ensures that the model's proficiency isn't skewed by imbalances or misconceptions across classes.

This robust CNN journey, intertwined with architectural nuances, training strategies, and evaluation frameworks, showcases the potency of deep learning in untangling the intricate nuances of musical instrument classifications. The harmonious interplay of layers, activations, and optimization techniques emerges as a testimony to the potential of CNNs in decoding the auditory world.

The journey of unraveling musical instrument classifications continued with the exploration of Artificial Neural Networks (ANNs), offering insights into their distinctive characteristics and their role in shaping the classification narrative. While the core architecture remains akin to the CNN, there exist nuances that warrant exploration.

In contrast to the CNN's affinity for 2D images, ANNs demand a distinct perspective – data presented in a 1D vector format (Shukla et al., 2020). This necessitated the transformation of the spectrogram images from their original 2D representation to a streamlined 1D format. This shift in perspective is vital, aligning with ANNs' expectation of a linear input structure.

Echoing the CNN journey, ANN's architectural framework embarks with a dense layer housing 64 units (Jeczmionek & Kowalski, 2021), fostering connectivity between neurons. The subsequent inclusion of a dropout layer at 0.5 magnitude underscores the commitment to averting overfitting, ensuring the model's generalization capabilities are fortified. (Srivastava et al., 2014)

The concluding layers mirror the orchestration within the CNN realm – a dense layer with 5 units representing each instrument class, complemented by a SoftMax activation function for multiclass classification. The essence of a learning rate set at 0.01 and the utilization of an Adam optimizer persist, fostering efficient convergence and adaptability.

It's worth noting that for the ANN model, a different approach was taken for hyperparameter tuning. Grid Search (Chowdhury et al. 2022), was employed to systematically explore various combinations of hyperparameters and identify the optimal set. This rigorous search process yielded a set of hyperparameters that significantly enhanced the model's performance. The hyperparameters obtained from Grid Search were then utilized to construct the ANN model.

The choice of categorical cross entropy as the loss function retains its significance in the ANN context, embodying its suitability for multiclass classification tasks. This loss function's utility extends to ANNs, preserving its role in calibrating predictive probabilities and actual class representations. (Jeczmionek & Kowalski, 2021)

Akin to the CNN journey, early stopping continues its crucial role in shaping the ANN's training process. Its vigilance over validation loss safeguards the model against overfitting, optimizing its prowess to encompass a wider spectrum beyond the training data. (Prechelt, 2002)

Training the ANN unfolds with familiar precision – a batch size of 32 and a sequence of ten epochs. The culmination yields an accuracy of 85%, underscoring the model's proficiency in classifying musical instruments. The evaluation journey journeys beyond accuracy, leaning on the foundational strength of the confusion matrix to refine the understanding of true positives, true negatives, false positives, and false negatives.

# Discussion

In the pursuit of classifying musical instruments based on spectrograms and audio features, a series of machine learning models were meticulously trained and evaluated. This tables presents a comparative overview of the models' performances, including Linear SVM, Polynomial SVM, RBF SVM, Random Forest, and Decision Tree. Each model was fine-tuned with optimized hyperparameters, and their accuracy, recall, precision, and F1-scores are analyzed across multiple instrument classes. This comprehensive assessment sheds light on the effectiveness of these models in the intricate task of musical instrument classification.

**TABLE 1:** Accuracy and Hyperparameters of Machine Learning Models

|  |  |  |
| --- | --- | --- |
| Model | Best Hyperparameters | Accuracy |
| Linear SVM | {'C': 1, 'kernel': 'linear'} | 61% |
| Polynomial SVM | {'C': 100, 'degree': 3, 'kernel': 'poly'} | 79% |
| RBF SVM | {'C': 10, 'gamma': 0.1, 'kernel': 'rbf'} | 81% |
| Random Forest | {'bootstrap': False, 'max\_depth': None, 'min\_samples\_leaf': 2, 'min\_samples\_split': 2, 'n\_estimators': 150} | 79% |
| Decision Tree | {'criterion': 'gini', 'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2} | 70% |

**TABLE 2**: Recall, Percision and F1-Score of Machine Learning Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Instruments | Recall | Percision | F1- Score |
| Linear SVM | Flute  Guitar acoustic  piano,  trumpet  violin | 0.57  0.68  0.52  0.64  0.61 | 0.6  0.66  0.58  0.52  0.68 | 0.58  0.67  0.55  0.58  0.64 |
| Polynomial SVM | Flute  Guitar acoustic  piano,  trumpet  violin | 0.80  0.81  0.72  0.86  0.74 | 0.72  0.82  0.79  0.82  0.77 | 0.76  0.81  0.75  0.84  0.75 |
| RBF SVM | Flute  Guitar acoustic  piano,  trumpet  violin | 0.74  0.87  0.79  0.79  0.81 | 0.72  0.84  0.90  0.82  0.71 | 0.73  0.85  0.84  0.80  0.76 |
| Random Forest | Flute  Guitar acoustic  piano,  trumpet  violin | 0.68  0.88  0.81  0.75  0.81 | 0.72  0.81  0.86  0.78  0.74 | 0.70  0.85  0.84  0.77  0.78 |
| Decision Tree | Flute  Guitar acoustic  piano,  trumpet  violin | 0.58  0.80  0.67  0.68  0.79 | 0.65  0.67  0.75  0.69  0.75 | 0.61  0.73  0.70  0.68  0.77 |

The linear SVM achieved an accuracy of 61%. It showed varying levels of recall, precision, and F1-score for different instrument classes. The model performed reasonably well but may benefit from further optimization.

The polynomial SVM with C=100 and a degree of 3 achieved an accuracy of 79%. It demonstrated higher recall, precision, and F1-scores compared to the linear SVM, indicating improved performance.

The RBF kernel SVM with C=10 and gamma=0.1 outperformed the previous models with an accuracy of 81%. It showed high recall, precision, and F1-scores, suggesting strong performance in classifying musical instruments.

The Random Forest model achieved an accuracy of 79%, on par with the polynomial SVM. It demonstrated good recall, precision, and F1-scores across instrument classes, making it a competitive choice for classification tasks.

The Decision Tree model had an accuracy of 0.70. While it had moderate recall and precision values, it showed room for improvement compared to the other models.

The RBF kernel SVM and the Random Forest model performed exceptionally well in classifying musical instruments based on spectrograms and audio features. These models demonstrated high accuracy and balanced recall, precision, and F1-scores across different instrument classes.

Certainly, the confusion matrix is a powerful tool to visually assess the performance of classification models (Racharla et al., 2020), and it can provide insights into how each class was classified. Let's discuss the confusion matrices for the SVM classifiers with different kernels.

In the confusion matrix for the Linear SVM, observe that it struggled to distinguish between certain musical instruments. There might be more misclassifications, especially among instruments with spectral similarities, like flute and piano. The diagonal elements of the confusion matrix (the true positives) would indicate correct classifications, while off-diagonal elements would represent misclassifications.

Figure 1: Confusion Matrix Linear SVM

The confusion matrix for the Polynomial SVM should show better performance compared to the Linear SVM. This kernel was able to capture some non-linear relationships in the data, resulting in improved accuracy. You would likely see fewer misclassifications, especially among instruments with more distinct acoustic characteristics.



Figure 2: Confusion Matrix Polynomial SVM

The confusion matrix for the RBF SVM, which achieved the highest accuracy among the SVM models, demonstrated even fewer misclassifications. This kernel excels at capturing complex, non-linear relationships in the data. would expect to see fewer off-diagonal elements, indicating more correct classifications.

Figure 3: Confusion Matrix RBF SVM

The Decision Tree, although interpretable, didn't match the accuracy of the RBF kernel SVM or the Random Forest. It performed reasonably well but struggled with instruments that shared spectral similarities.

The Random Forest, as an ensemble of decision trees, showcased both interpretability and accuracy. Its aggregated decisions and feature importance assessment made it a strong performer in this classification task.



Figure 4: Confusion Matrix Decision Tree



Figure 5: Confusion Matrix Random Forest

By visually inspecting the confusion matrices for these SVM kernels, you can get a clearer picture of their performance and see where they excel or struggle. The goal is to minimize the off-diagonal elements, especially in the rows and columns corresponding to instruments that are more acoustically similar.

The following table summarizes the best hyperparameters and model accuracy for both the ANN and CNN models:

|  |  |  |
| --- | --- | --- |
| Model | Hyperparameters | Accuracy |
| ANN | {'units1': 100, 'units2': 50, 'units3': 25, 'learning\_rate': 0.001} | 85% |
| CNN | {'units1': 25, 'units2': 50, 'units3': 100, 'learning\_rate': 0.001} (Manual Tuning) | 99% |

**Table 3**: Parameters for Neural Network Models

The ANN model achieved an accuracy of 85%, which is a respectable performance, suggesting that the hyperparameters used were effective for this task.

The CNN model, on the other hand, achieved an impressive accuracy of 99% after manual hyperparameter tuning. This exceptional accuracy highlights the potential of CNNs for extracting complex features from spectrogram data, resulting in highly accurate classifications. While both models performed well, the CNN model, with manual hyperparameter tuning, exhibited outstanding accuracy, making it a robust choice for the classification of musical instruments based on spectrogram data.

The learning curves, along with the trends in training and testing accuracy and loss, provide valuable insights into how well your Convolutional Neural Network (CNN) and Artificial Neural Network (ANN) models were trained to classify musical instruments based on spectrogram data. Here's a summary of observations

Both training and testing accuracy on the CNN are almost identical, indicating that the CNN model performed consistently well on both the data it was trained on and unseen data. The loss on the CNN model also follows a similar trend for both training and testing data, showing that the model was able to fit the data effectively without overfitting.



Figure 6: Learning Curve CNN

The ANN model's training and testing accuracy also showed similar performance, suggesting good generalization. However, towards the end of training, a slight difference between training and testing accuracy emerged. Similarly, the loss on the ANN model started off well, but towards the end of training, a slight divergence between training and testing loss occurred, aligning with the accuracy trends.

****

Figure 7: Learning Curve ANN

The CNN model appears to have trained very well, with consistent accuracy and loss trends, indicating effective learning and generalization.

The ANN model performed reasonably well, but the slight divergence between training and testing accuracy and loss towards the end of training.

Overall, the ROC curves and the trends in accuracy and loss provide a comprehensive view of how both models learned to classify musical instruments. While both models achieved good performance, the CNN model seems to have performed slightly better in terms of consistency between training and testing phases.

# CONCLUSION

The research question sought to determine the optimal combination of spectrogram and audio features, as well as the choice of deep learning and machine learning models, for achieving the highest accuracy in classifying various musical instruments based on spectrograms and audio features. Spectrograms and MFCCs were identified as valuable features for this task. Spectrograms provide a visual representation of audio data over time, capturing spectral patterns. MFCCs offer a numerical representation, distilling complex audio into concise coefficients. The combination of these features enriched the analysis. Convolutional Neural Networks (CNNs) emerged as the top-performing models in this classification task. Their ability to process spectrogram data, capturing spatial relationships, and intricate patterns proved highly effective. The CNN model achieved an impressive accuracy of 99%. Support Vector Machine (SVM) models, particularly with radial basis function (RBF) kernels, and Random Forest demonstrated competitive performance among traditional machine learning techniques. They achieved accuracies of 81%, showcasing their effectiveness. The project underscored the importance of data quality, including stereo audio conversion and standardized formats, in achieving accurate classification. Additionally, the choice of features and model architecture played pivotal roles in overall model performance. In conclusion, this research question led to valuable insights into the classification of musical instruments based on spectrograms and audio features. The study showcased the dominance of CNNs in this domain while also highlighting the competitiveness of traditional machine learning models. These findings offer a foundation for future work in music classification and information retrieval systems

# Future Research Directions

Future research could explore ensemble techniques, combining the strengths of deep learning models like CNNs with traditional models, to further enhance classification accuracy and robustness. Investigating the impact of data augmentation and larger datasets on model generalization would be beneficial. The findings provide valuable insights for advancing music classification systems, contributing to music recommendation, genre classification, and broader music information retrieval applications.

**REFERENCE**

1. Racharla, K., Kumar, V., Jayant, C. B., Khairkar, A., & Harish, P. Predominant Musical Instrument Classification based on Spectral Features. Indian Institute of Technology, Kharagpur India.
2. Deng, D., Simmermacher, C., & Cranefield, S. A Study on Feature Analysis for Musical Instrument Classification. Dept. of Information Science, University of Otago.
3. Solanki, A., & Pandey, S. (2019). Music instrument recognition using deep convolutional neural networks. 1 • Sachin Pandey.
4. Heittola, T., Klapuri, A., & Virtanen, T. Musical instrument recognition in polyphonic audio using source-filter model for sound separation.
5. Li, P., Qian, J., & Wang, T. Automatic instrument recognition in polyphonic music using convolutional neural networks.
6. Nam, J., Choi, K., Lee, J., Chou, S.-Y., & Yang, Y.-H. Deep Learning for Audio-Based Music Classification and Tagging.
7. Park, T., & Lee, T. (2015). Musical instrument sound classification with deep convolutional neural network using feature fusion approach.
8. Gururani, S., Summers, C., & Lerch, A. Instrument activity detection in polyphonic music using deep neural networks.
9. Tanaka, K., Nakatsuka, T., Nishikimi, R., Yoshii, K., & Morishima, S. (2020). Multi-Instrument Music Transcription Based on Deep Spherical Clustering of Spectrograms and Pitchgrams.
10. Gómez, J. S., Abeßer, J., & Cano, E. (2018). Jazz Solo Instrument Classification with Convolutional Neural Networks, Source Separation, and Transfer Learning. Proceedings of the 19th ISMIR Conference.
11. Chen, H.-H., & Lerch, A. (2022). Music Instrument Classification Reprogrammed. Music Informatics Group, Georgia Institute of Technology.
12. Ashraf, M., Abid, F., Ud Din, I., Rasheed, J., Yesiltepe, M., Yeo, S. F., & Ersoy, M. T. (2022). A Hybrid CNN and RNN Variant Model for Music Classification
13. Shukla, U., Tiwari, U., Chawla, V., & Tiwari, S. (2020). Instrument Classification using image-based Transfer Learning.
14. H., Zhang, Y., & Zhang, Q. The Use of Deep Learning-Based Intelligent Music Signal Identification and Generation Technology in National Music Teaching. Frontiers in Psychology
15. Li, P., Qian, J., & Wang, T. (2015). Automatic Instrument Recognition in Polyphonic Music Using Convolutional Neural Networks.
16. Kostek, B. (2004). Musical instrument classification and duet analysis employing music information retrieval techniques.
17. Elbir, Ahmet, et al, 2018, "Music Genre Classification and Recommendation by Using Machine Learning Techniques
18. Yang, Rui, et al., 2020, "Parallel recurrent convolutional neural networks-based music genre classification method for mobile devices.
19. McFee, Brian, et al, 2015, "librosa: Audio and music signal analysis in python
20. Corrêaa, Débora C. and Francisco Ap Rodriguesa., 2018, "A Survey of Symbolic-Based Music Genre Classification
21. Weninger F, Schuller B. Audio recognition in the wild: static and dynamic  
    classification on a real-world database of animal vocalizations.
22. Choi K, Fazekas G, Sandler M, Cho K. Transfer learning for music classification  
    and regression tasks
23. Cotton CV, Ellis DPW. Spectral vs. spectro-temporal features for acoustic event  
    detection
24. Khamparia A, Gupta D, Nguyen NG, Khanna A, Pandey B, Tiwari P. Sound  
    classification using convolutional neural network and tensor deep stacking  
    network.
25. Salamon J, Bello JP. Deep convolutional neural networks and data  
    augmentation for environmental sound classification.
26. Zhu, B., Wang, Y., Patel, R., & Li, Y. (2018). Learning environmental sounds with multi-scale convolutional neural network. In Proceedings of the International Joint Conference on Neural Networks.
27. Geiger, J.T., & Helwani, K. (2015). Improving event detection for audio surveillance using Gabor filterbank features.
28. Saldanha EL, Corso JF. Timbre cues and the identification of musical instruments. J. Acoust. Soc. Am. 1964
29. Heittola, T., Klapuri, A., & Virtanen, T. (Year). Musical instrument recognition in polyphonic audio using source-filter model for sound separation.
30. Han, Y., Kim, J., & Lee, K. (2017). Deep convolutional neural networks for predominant instrument recognition in polyphonic music.
31. Pons, J., Slizovskaia, O., Gong, R., Gómez, E., & Serra, X. (2017). Timbre analysis of music audio signals with convolutional neural networks.
32. Bhalke, D.G., Rao, C.B. Rama, & Bormane, D.S. (2016). Automatic musical instrument classification using fractional Fourier transform based-MFCC features and counter propagation neural network.
33. Li, J., Han, L., Li, X., Zhu, J., Yuan, B., & Gou, Z. (2022). An evaluation of deep neural network models for music classification using spectrograms. Multimedia Tools and Applications
34. Quazi, G.R., Mohammed, N., Sadia, Z.P., & Sabrina, A. (2021). Comparative analysis of three improved deep learning architectures for music genre classification.
35. Bhat, A.S., Amith, V.S., Prasad, N.S., & Mohan, D.M. (2014). An efficient classification algorithm for music mood detection in western and Hindi music using audio feature extraction.
36. Li, T., Ogihara, M., & Li, Q. (2003). A comparative study on content-based music genre classification
37. Jensen, K. (1999). Pitch independent prototyping of musical sounds
38. Muhammad, G., & Alghathbar, K. (2009). Environment recognition from audio using MPEG-7 features
39. Baniya, B.K., Lee, J., & Li, Z.N. (2014). Audio feature reduction and analysis for automatic music genre classification.
40. Scheirer, E.D. (1998). Tempo and beat analysis of acoustic musical signals.
41. Zheng, E., Moh, M., & Moh, T.-S. (2017). Music genre classification: A n-gram based musicological approach.
42. Sigtia, S., & Dixon, S. (2014). Improved music feature learning with deep neural networks.
43. Wang, Z., Muknahallipatna, S., Fan, M., Okray, A., & Lan, C. (2019). Music Classification using an Improved CRNN with Multi-Directional Spatial Dependencies in Both Time and Frequency Dimensions.
44. Maccagno, A., Mastropietro, A., Mazziotta, U., Scarpiniti, M., Lee, Y., & Uncini, A. (2019). A CNN Approach for Audio Classification in Construction Sites.
45. Bosch, J.J., Janer, J., Fuhrmann, F., & Herrera, P. (2012). A Comparison of Sound Segregation Techniques for Predominant Instrument Recognition in Musical Audio Signals.
46. Piczak, K.J. (2015). Environmental sound classification with convolutional neural networks. In Machine Learning for Signal Processing (MLSP)
47. Rajesh, S., & Nalini, N.J. (2020). Musical instrument emotion recognition using deep recurrent neural network
48. Elsayed, G.F., Goodfellow, I., & Sohl-Dickstein, J. (2019). Adversarial reprogramming of neural networks.
49. Gong, Y., Chung, Y.A., & Glass, J.R. (2021). AST: Audio spectrogram transformer. In Interspeech.
50. Gururani, S., Lerch, A.: Semi-supervised audio classification with partially labeled
51. Gururani, S., & Lerch, A. (2021). Semi-supervised audio classification with partially labeled data.
52. Bhalke, D., Rao, C., & Bormane, D.S. (2016). Automatic musical instrument classification using fractional Fourier transform based-MFCC features and counter propagation neural network. Journal of Intelligent Information Systems.
53. Ye, & Yang, J. (2021). A deep neural network model for speaker identification. Applied Sciences.
54. Naranjo-Alcazar, J., Perez-Castanos, S., Zuccarello, P., Antonacci, F., & Cobos, M. (2020). Open set audio classification using autoencoders trained on few data. Sensors.
55. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation.
56. Choi, K., Fazekas, G., Cho, K., Sandler, M., & Cunningham, S. (2017). A tutorial on deep learning for music information retrieval
57. Wu, W., Xia, Z., Qian, Y., Li, Z., & Li, Z. (2018). Music genre classification using independent recurrent neural network.
58. Fulzele, P., Bhadane, M., Kasar, T., & Kulkarni, S. (2018). A hybrid model for music genre classification using LSTM and SVM.
59. Duan, Z., Han, J., & Pardo, B. (2014). Multi-pitch streaming of harmonic sound mixtures. IEEE/ACM Transactions on Audio, Speech, and Language Processing.
60. Zahid, S., Hussain, F., Rashid, M., Yousaf, M. H., & Habib, H. A. (2015). Optimized Audio Classification and Segmentation Algorithm by Using Ensemble Methods.
61. Rodriguez, A., 2022. Deep Learning Systems: Algorithms, Compilers, and Processors for Large-Scale Production. Springer Nature.
62. Dalyac, A., Shanahan, M. and Kelly, J., 2014. Tackling class imbalance with deep convolutional neural networks. *Imperial College*, *2*, pp.30-35.
63. Safavian, S.R. and Landgrebe, D., 1991. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, *21*(3), pp.660-674.
64. Afram, A. and Sarab Fard Sabet, D., 2023. Exploring State-of-the-Art Machine Learning Methods for Quantifying Exercise-induced Muscle Fatigue.
65. Haykin, S. 2009. Neural Networks a Comprehensive Foundation, 8th Edn., Pearson Prentice Hall, India.
66. Whibley, S. (2016). WAV Format Preservation Assessment (Version 1.0) [External].
67. AltexSoft & Bruxlab. (2022, May 12). Audio Analysis with Machine Learning: Building AI-Fueled Sound Detection App. AltexSoft. URL: <https://www.altexsoft.com/blog/audio-analysis/>
68. Juan J. Bosch, Ferdinand Fuhrmann, and Perfecto Herrera (2012) IRMAS: A Dataset for Instrument Recognition in Musical Audio Signals. Zenodo
69. Vakili, M., Ghamsari, M. and Rezaei, M., 2020. Performance analysis and comparison of machine and deep learning algorithms for IoT data classification
70. Gopal, S., & Yang, Y. (2014). Von Mises-Fisher Clustering Models. In International Conference on Machine Learning (ICML) (Vol. 1, pp. 269–302).
71. Jia, X. (2022). A Music Emotion Classification Model Based on the Improved Convolutional Neural Network. Computational Intelligence and Neuroscience, 2022, Article ID 6749622, 11 pages. <https://doi.org/10.1155/2022/6749622>
72. Parker, C., Scott, S. and Geddes, A., 2019. Snowball sampling. *SAGE research methods foundations*.
73. GDPR.eu. (2023). Article 6
74. Bartusiak, E.R. and Delp, E.J., 2022. Frequency Domain-Based Detection of Generated Audio. VIPER Laboratory, School of Electrical and Computer Engineering, Purdue University; West Lafayette, IN, USA.
75. iZotope. (2020). Understanding Spectrograms. Retrieved from <https://www.izotope.com/en/learn/understanding-spectrograms.html>
76. Yang, G., Liu, W., Liu, X., Gu, X., Cao, J., & Li, J. (2020). Delving into the Frequency: Temporally Consistent Human Motion Transfer in the Fourier Space
77. IBM. (2023). Random Forest. Retrieved from <https://www.ibm.com/topics/random-forest>
78. Agrawal, N. (2023). Decoding the Symphony of Sound: Audio Signal Processing for Musical Engineering.
79. O’Shea, K., & Nash, R. (2015). An Introduction to Convolutional Neural Networks.
80. Deng, L., 2012. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6), pp.141–142.
81. Yang, L., & Shami, A. (2022). On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice.
82. Veeralagan, J., & Manju Priya, S. (2022). Hyper Tuning Using Gridsearchcv on Machine Learning Models for Prognosticating Dementia.
83. Prabavathy, S., Rathikarani, V., & Dhanalakshmi, P. (2020). Classification of Musical Instruments using SVM and KNN.
84. Kour, G., & Mehan, N. (2015). Music Genre Classification using MFCC, SVM, and BPNN
85. Rajesh, S., & Nalini, N. J. (2019). Musical instrument emotion recognition using deep recurrent neural network.
86. Bergstra, J., Yamins, D., & Cox, D. D. (2013). Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures.
87. Chowdhury, A.A., Das, A., Hoque, K.K.S. and Karmaker, D., 2022, May. A Comparative Study of Hyperparameter Optimization Techniques for Deep Learning. In *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2021* (pp. 509-521). Singapore: Springer Nature Singapore.
88. Ahsan, M. M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance.
89. Rodríguez, P., Bautista, M. A., González, J., & Escalera, S. (2018). Beyond One-hot Encoding: lower dimensional target embedding.
90. Sakib, S., Ahmed, N., Kabir, A. J., & Ahmed, H. (2019). An Overview of Convolutional Neural Network: Its Architecture and Applications.
91. Jeczmionek, E., & Kowalski, P. A. (2021). Flattening Layer Pruning in Convolutional Neural Networks.
92. Prechelt, L. (2002). Early stopping-but when? In Neural Networks: Tricks of the Trade (pp. 55-69). Berlin, Heidelberg: Springer Berlin Heidelberg.
93. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Department of Computer Science, University of Toronto, 10 Kings College Road, Rm 3302, Toronto, Ontario, M5S 3G4, Canada.