Section 2

A) Review of research methodologies & map

Categorization of musical instrument using machine learning has attracted a lot of interest due to its use in a wide variety of applications such as smart music systems, musical analysis and audio search engines. Despite achievements and continuous progress in the field, it is still a complex problem to differentiate between some instruments that possess similar characteristics like timbral and pitch properties, such as the flute and the clarinet. To assist and contribute to the ever-growing knowledge of this field, researchers have utilized a variety of deep learning techniques, mainly involving Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid CNN-RNN architectures.  
This section reviews key method, methodological approaches adopted in the related studies, focusing on three aspects: the preprocessing, the overall structure of the learning architecture, as well as the evaluation method. Figure 2 below shows a visual representation of the chosen literature, providing a bigger understanding of the main contributions to the study and any connections between the literature.

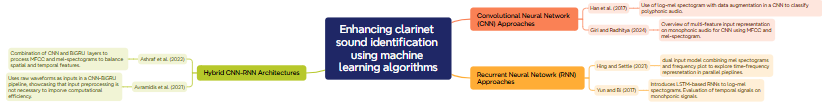


Fig. 2. The Literature Map

1. Convolutional Neural Network (CNN) Approaches

The use of convolutional neural networks (CNNs) in musical instrument identification have been utilized in several studies. As discussed in [1], the study aimed to use spectrogram images as an input for the model to identify instruments in a polyphonic environment. Before transforming the audio dataset into log-mel spectrograms, several data augmentation techniques such as pitch shifting, setting a fixed length of the audio clips and removing background noise were applied to enhance the dataset and mirror real-life music conditions. The CNN structure consists of three convolutional layers, with ReLu activation functions, each followed by max-pooling layers. The output was flattened and forwarded to two fully connected layers and a SoftMax classifier which transforms raw outputs into probabilities. To evaluate the findings, cross-validation was employed and drawbacks in the study such as class imbalances of the instruments were handled by metrics such as F1-Score and precision.

In contrast to [1], the researchers of [2] focused on the classification of monophonic audio signals of four different instruments instead of polyphonic music by using different configurations of a CNN based on frequency-based features. A variety of feature extraction techniques were applied, they calculated 13 coefficients of MFCC as well as Mel-Spectrograms like [1], the techniques were normalized and merged into a singular multidimensional array. The implementation consists of a two layer CNN, with each having a convolutional kernel with ReLu activations and max pooling layer. To reduce overfitting, a global average pooling (GAP) layer was used to reduce flattening. Similarly to [1], a SoftMax output layer is utilized. Additional experimentation was planned to overview differences between feature extraction techniques by separating the feature extraction techniques and running the CNN model on a singular feature extraction technique.

The researchers of [3], built on the study of [2], by incorporating frequency-based input features and also takes account temporal recurrence by using dual-image inputs. This technique was done by combining the Mel-Spectrograms and recurrence plots of the NSynth dataset. The Mel-Spectrograms were sampled to 1024 windows and the recurrence plots were created with a phase space embedding approach. The images were concatenated along the channel axes to generate a dual-input tensor. For the inputs, two parallel designed CNN pipelines were constructed, with each stream consisting of two convolutional blocks and max pooling layer. The input for the dense layer was the concatenation of the output of both streams. The evaluation process was similar to [1] [2], using a SoftMax output node to report precision and recall of the classes.

1. Recurrent Neural Network (RNN) Approaches

Experimentation of instrument identification through the use of Recurrent Neural Networks (RNNs) is less common than CNNs. Despite that, they offer alternative advantages when using specific features such as sequential features or time-dependent features.

The researchers in [4] investigated and compared both CNN and Long Short-Term Memory (LSTM)-based RNN to classify musical instrument recordings of a monophonic environment. The recordings were re-sampled to 22.05 kHz and normalized to set a consistent amplitude range. Similarly to previous studies, the input feature was a log-scale Mel-Spectrograms. These spectrograms were suited to the LSTM model as they retained temporal evolution of the frames. This approach took a different turn from the traditional image-based CNN. As the spectrograms were transformed into time-step frequencies, with each frame representing a vector of frequency magnitudes.

The model implemented in [4] consisted of a single-layer LSTM using a dropout layer between the time steps to avoid overfitting. An input of a 200-time step sequence was used with each having 128 frequency components. The output of the LSTM was passed to a fully connected layer, also with a softmax activation to classify the instruments. The parallel CNN model was trained with the same dataset and features in order to ensure consistency. Both models were trained on the Adam optimizer, which included early stopping based on validation accuracy. The evaluation phase consisted of metrics similar to previous study for instrument classification such as F1-Score and accuracy.

1. Hybrid CNN-RNN Architectures

Hybrid CNN-RNN Architectures involve the combination of the CNNs abilities to capture spatial feature extractions as well as the RNNs strength in capturing the temporal sequence. Certain tasks that involve audio signals can benefit from this approach as both frequency patterns and temporal signals are important for a more accurate classification. In [5] and [6], we see different implementations of this hybrid approach.

In [5], Ashraf et al. proposed the use of the traditional input features such as MFCC and Mel-Spectrograms for the CNN-RNN hybrid model. The implementation pipeline started with a Fast Fourier Transform (FFT) to sample audio files to 16 kHz mono each and generate Mel-spectrograms. Alongside this, delta features were used to extract 40-dimensional MFCCs, creating a feature matrix that incorporated both dynamic and timbral features. The researchers accounted for stabilized input dimensions, thus all feature matrices were standardized and split into non-overlapping intervals of three seconds each.

The hybrid architecture was structured to use two 2D convolutional layers followed by a max pooling layer. Given the nature of the hybrid model, a Bidirectional Gated Recurrent Unit (BiGRU) was used to process temporal sequences in a bi-directional way, capturing long-range dependencies from the resulting feature maps that were reshaped into sequential vectors. Training was performed by the Adam optimizer and evaluation was done using a 10-fold cross-validation.

The study of [6] disregard the conservative input features like the Mel-spectrogram and MFCC and adopt an alternative approach by employing an end-to-end technique using the raw monophonic signals as direct input. These are signals were normalized in a 16kHz sample rate and passed through a series of 1D convolutional layer and reshaped the input into a sequence of feature vectors, opposed to the sequential vectors of [5]. These vectors were passed to a BiGRU layer, and a final dense layer and a softmax classifier generated the instrument predictions and classification.  
Unlike [5], the methodology of [6] was less complex due to the straightforward preprocessing pipeline it required, emphasizing an ideal alternative approach with an emphasis on computational efficiency.

B) Methodology

3.1 Research Questions

The main focus of this study is to explore the effectiveness of different machine learning architectures in the successful identification of clarinet sounds from audio samples. This study aims to address two research questions, which are stated as follows:

* **RQ1 – Which machine learning algorithm between the CNN, RNN and hybrid CNN-RNN is best at identifying clarinet sounds within a polyphonic environment?**

With the use of consistent input data and training settings, this research question aims to evaluate and compare and contrast each algorithm’s classification performance with one another. Metrics such as accuracy, and F1-Score will be employed to show each models effectiveness. The main point will be how good each model can differentiate the clarinet from other instruments with similar pitch and timbre such as flute, and oboe.

* **RQ2 – How do audio feature extraction techniques affect the overall performance of machine learning models in clarinet sound identification?**

This research question aims to explore the effect, input representations leave on the classification outcomes. This question will help identify which features best target the unique characteristics and harmonics of the clarinet and contribute to improved accuracy. All feature extraction techniques are outlined in Table 1.

Table 1: Feature Extraction Techniques

|  |  |  |
| --- | --- | --- |
| **Feature Type** | **Description** | **What it Captures** |
| Mel-Spectrogram | Time-frequency representation with perceptual frequency. | Frequency distribution across time |
| MFCC (Mel Frequency Cepstral Coefficients) | It models how humans perceive sound suing cepstral analysis | Timbre and perceptual tone qualities |
| Chroma Features | Represents energy in each of the 12 different pitch class of music | Harmonic and Pitch Content |

3.2 Research Design

3.2.1 Research Objectives

The main objective of this study is to develop and evaluate three different machine learning models that can consistently identify clarinet sounds in a polyphonic environment. The chosen architectures to be created and trained are: Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and a hybrid architecture (CNN-RNN). To achieve a uniformed structure, the IRMAS dataset [7], a widely known dataset in the field of musical instrument categorization, will be used to train all the models. The models used are outlined in the Table 2 below.

Table 2: Model Architectures Overview

|  |  |  |
| --- | --- | --- |
| **Model** | **Core Components** | **Primary Focus** |
| CNN | Convolutional layers, Max pooling, Dense layer | CNN extracts spatial features from time-frequency images |
| RNN (LSTM) | LSTM layers, Dense layers | Captures sequential and temporal patterns |
| Hybrid CNN-RNN | Convolutional layers, BiGRU (Bidirectional GRU), Dense layer | Combines spatial features and sequential modelling |

Another objective of this research is to investigate the different feature extraction techniques such as: Mel-Spectrogram, MFCCs, and Chroma Features effect the classification performance. The chosen representations each capture different qualities of musical audio like the timbral, spectral, and harmonic qualities, which are necessary to distinguish the clarinet from other similar instruments.

To ensure a fair and consistent comparison, the study intends to present a strong experimental framework that integrates, the dataset split, input dimensions, and evaluation metrics across all the models. The study will give importance the incorrect classification patterns and the overall abilities of each model to identify clarinet sounds. The results that will be obtained will provide important information on which model and audio representations are best suited for further problems in musical instrument classification.

3.2.2 Justification of Chosen Methodology

The use of machine learning-based experimental methodology is justified by the nature of the research objectives, which aim to calculate the efficiency of different algorithm techniques in musical instrument identification. The analysis of audio signals represented by the feature inputs mentioned, creates a perfect fit for deep learning models, specifically for CNN, which is primarily good at extracting spatial features, RNN, is suited to capture sequential data, and hybrid CNN-RNN architectures which is made to use both spatial and temporal data at the same time. These models have previously shown very good results in tasks which involve time-series or image-like data.

To understand how various audio signal representations effect the performance of each model, a variety of feature extraction techniques were included. Each feature captures a distinct aspect of musical sounds: Mel-Spectrograms capture frequency distribution, MFCC represents human auditory perception, and Chroma Features highlight harmonic content, all which can improve the model’s reaction to input data which may lead to better results.

All things considered, the chosen methodology aligns well with the goal of the study to provide a thorough, data-driven evaluation of machine learning algorithms for identifying clarinet sounds, which guarantees validation and a practical application.

3.3 Methodology Approach

As part of the experimental design, three different models will be developed and trained: CNN, RNN and a hybrid CNN-RNN Architecture. To keep uniformity and consistent results, the same dataset will be applied to each model, as well as the same feature extraction techniques and training settings. To also assess the effect of input representations on classification results, the study will employ a variety of feature extraction approaches, including Mel-Spectrograms, MFCCs, and Chroma features.

The dataset that will be used for this experimentation is the IRMAS dataset (Instrument Recognition in Musical Audio Signals) [7]. The dataset contains annotated recordings of 11 musical instruments. The length of each audio clip is set between 3 to 15 seconds, and each clip is labelled according to the predominant instrument. The main reason for this choice of dataset is thanks to its real-world audio conditions such as background noise and polyphonic form. The clarinet audio samples will be the focus of the study, along a selection of instrument with similar characteristics like the flute, saxophone, and oboe. This will guarantee that the classification task is related to the goal of the study.

Certain preprocessing steps will be done when required. The audio clips will be resampled to 22.05 kHz, converting them from stereo to mono recordings, and the duration of the clips will be cut to set a uniform duration. The dataset will then be divided into a 70 – 15 – 15 formats of training, validation and testing subsets.

Analysis of each model’s performance will use standard classification metrics including accuracy, precision, recall, and F1-Score. A comparison between these metrics will determine which feature extraction method best captures the clarinet’s qualities as well as which machine learning architecture delivers the best overall performance.

3.4 Validity, Reliability, and Generalizability

Many parameters and steps were considered to ensure experimentation validity. First, all the models (CNN, RNN, and hybrid CNN-RNN) are trained and evaluated under similar conditions, using the same dataset and dataset split, feature extraction techniques, as well as hyperparameter settings. Finally, to ensure that the differences seen are really due to the model architecture or the feature extraction, the study mitigates any confounding factors that may affect model performance by keeping a consistent experimental framework.

The reliability of the experimentation was ensured by using a consistent and reproducible procedure. The experimentation can be repeated under the same conditions as the dataset split is fixed. Additionally, other researchers are able to repeat the experimentation without depending on privately owned technology as it was done on publicly available tools and libraries such as Keras and LibROSA. Cross-validation also help the reliability of the model and prevent random fluctuations during the training process from influencing the results.

Several factors must be acknowledged when evaluating the findings of the study in terms of its generalizability and transferability. The IRMAS dataset supports generalizability since it comprises of a variety of genres and real-world recording instances [7]. On the other hand, although the focus of the study is identification of the clarinet, the data gathered from this study might be useful to other general musical instrument classification tasks as the dataset allows for the inclusion of instruments with similar characteristics.

3.5 Ethical Considerations

To maintain integrity and transparency, a set of ethical principles were followed during the period of this research. Firstly, there was no direct contact with any human, as well as not any private information and private data were utilized in this research. Secondly, the IRMAS dataset is free to use for academic purposes and for the protection of intellectual property the dataset is cited and acknowledged [7].

Additionally, the software and libraries are also open source, minimizing difficulties associated with access inequality. To promote research procedures, the conversion of data, feature extraction, and experimental models have all been recorded and mentioned in the study. Overall, this study satisfies the ethical requirements in the fields of musical analysis and machine learning.

References

[1] Y. Han, J. Kim, and K. Lee, "Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 208-221, Jan. 2017. [Online]. Available: <https://ieeexplore.ieee.org/document/7755799>

[2] K. Giri and H. Radhitya, “Musical instrument classification using audio features and convolutional neural network,” in *Proc. International Conference on Informatics*, 2024.

[3] S. Hing and A. Settle, “Detecting and classifying musical instruments with convolutional neural networks,” *CS230 Stanford Projects*, Winter 2021. [Online]. Available: <https://cs230.stanford.edu/projects_winter_2021/reports/70770755.pdf>

[4] M. Yun and J. Bi, “Deep learning for musical instrument recognition,” University of Rochester, 2017. [Online]. Available: <https://hajim.rochester.edu/ece/sites/zduan/teaching/ece477/projects/2017/MingqingYun_JingBi_ReportFinal.pdf>

[5] M. Ashraf, M. Rahman, and R. Hasan, “A hybrid CNN and RNN variant model for music classification,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 13, no. 12, pp. 237–244, 2022. [Online]. Available: <https://www.researchgate.net/publication/367343673_A_Hybrid_CNN_and_RNN_Variant_Model_for_Music_Classification>

[6] V. Avramidis, A. Pantelopoulos, and E. Kaimakamis, “Deep convolutional and recurrent networks for polyphonic instrument classification from monophonic raw audio waveforms,” *arXiv preprint arXiv:2102.06930*, 2021. [Online]. Available: <https://arxiv.org/abs/2102.06930>

[7] [7] J.J. Bosch, J. Janer, F. Fuhrmann, and P. Herrera, "A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals," in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, Porto, Portugal, 2012, pp. 559–564.