Section 4 – Results, analysis and discussion

As discussed in section 3, the identification of the clarinet against other instruments with similar characteristics was evaluated through the three different experiments. The audio features were extracted using several techniques such as Mel-Spectrograms, MFCCs, and chroma features depending on the specific architecture of each model – CNN, RNN, and CRNN respectively. To digest the obtained results, metrics like accuracy, precision and f1-score were used to provide a macro-level overview of classification accuracy and a micro-level insight into each model’s ability to accurately organise each instrument class. Since the focus of the study revolves around the clarinet, more focus will be put on how each model accurately classifies the clarinet specifically. The model’s performance results accompanied by confusion matrices and other important graphs, and an explanation of the findings as well as a comparison between each result are discussed in the following subsections.

A fixed set of evaluation metrics were used to evaluate each model’s performance: accuracy, precision, F1-score, and recall. To guarantee that every instrument class, were given equal weights, the metrics were calculated on a per-class basis and averaged using macro averaging. Below the metrics are detailed in table 1.

|  |  |
| --- | --- |
| Metric | Description |
| Accuracy | Overall correctness of the model over each class |
| Precision | Shows how often a predicted instrument label is correct. |
| F1-Score | Balanced metrics between precision and recall, useful for equal class distributions |
| Recall | Measures true positives that were identified, showing how well the model captures true occurrences |

In addition to these metrics, confusion matrices were also created to visualize misclassifications between the instrument classes and spot any relevant trends in instrument confusion.

CNN Model Results and Discussion

The CNN model was the first architecture in the study, utilizing the mel-spectrograms as feature input representation. The model achieved a moderate performance overall among the four instrument classes (clarinet, trumpet, saxophone, and flute), achieving a validation accuracy of 57.41%.

For the clarinet categorization, the CNN achieved a precision of 0.52, an F1-Score of 0.44, and a recall of 0.38. The lower recall shows that the model missed a substantial amount of real clarinet instances, even if the accuracy result suggests that over half of the predicted “clarinet” labels were accurate. These relatively low metrics can be explained by the CNNs incapacity to detect temporal features, an important feature for instrument classification since fluctuations of instruments timbre over time are important.

The confusion matrix (Figure 4.1) provides further data on the model’s errors. Out of the 101 clarinet samples, only 38 were accurately classified. The remaining samples were misclassified as: Flute (20), Saxophone (23), Trumpet (20). This suggests that the CNN particularly struggles to differentiate the tone and harmonic profile of the clarinet from the other instruments.

A diagram of a confusion matrix

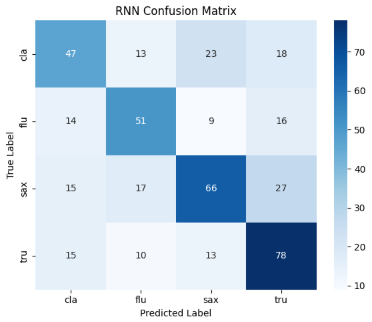
AI-generated content may be incorrect.

RNN Model Results and Discussion

The RNN architecture, focuses on capturing temporal and sequential audio data is combined with MFCC as feature input. Since RNNs can gather data over time steps, it makes it more suited for time-series data than CNNs. With an overall validation accuracy of 56.06% across all the classes, the RNN model performed a bit worse than the CNN. However, on the clarinet class, it performed reasonably better.

For the clarinet class, the RNN architecture achieved a precision of 0.47, a recall of 0.49, and an F1-Score of 0.48. Despite a lower precision than the CNN, the RNN greatly improves in the recall (+0.11) suggesting that it identified more instances of the clarinet.

The confusion matrix (Figure 4.2) shows how the clarinet samples were interpreted. Out of the same 101 samples, these were how they were classified: Clarinet (47), Flute (13), Saxophone (23), Trumpet (18). Compared to the CNN, which only produces 38 valid predictions, the RNN demonstrates a reduction in clarinet misclassifications. Despite the saxophone still causing the most confusion, the temporal modelling of the RNN seems to help in capturing the required tonal characteristics that define the clarinet.



CRNN Model Results and Discussion

The CRNN model was designed to combine the strengths of both CNN and RNN. This hybrid model was trained using chroma features, emphasizing on harmonic and pitch context. The CRNN achieved an overall validation accuracy of 61.34%, outperforming both the CNN and RNN models.

The CRNN model’s performance indicates a more balanced improvement across all metrics. Focusing on the clarinet metrics, the hybrid architecture achieved a precision of 0.52, recall of 0.50, and F1-Score of 0.51. The increase in F1-Score shows that the CRNN provide the best percentage between true positives and false positives. Furthermore, the obtained recall indicates that this model was better at recognizing clarinet samples whilst keeping a low misclassification rate.

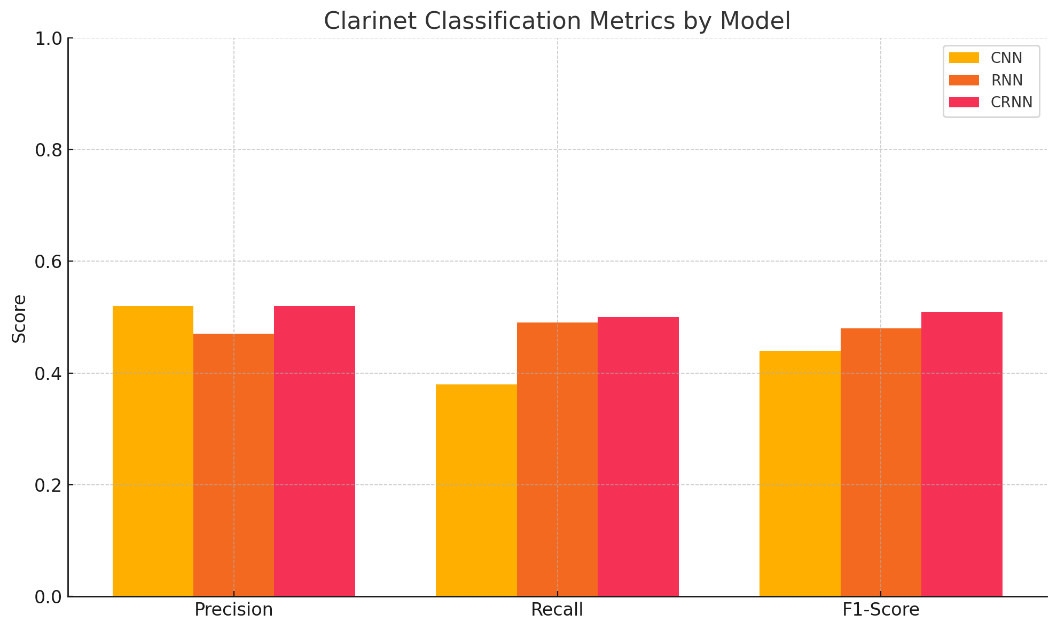
According to the confusion matrix (Figure 4.3) seen below, 50 samples were correctly classified as clarinet, with the other misclassifications: Flute (19), Saxophone (24), Trumpet (10). Out of all the models, the CRNN provided the highest correct number of clarinet correct predictions, suggesting that combining spatial and temporal modelling works well for this classification. This finding supports the hypothesis that hybrid systems provide improvements in audio-based classification performance, particularly in slight differences between similarly sounding woodwind instruments and brass instruments.

A diagram of a graph

AI-generated content may be incorrect.

Comparative Analysis of All Models

This section discusses each model independently and then combines the results for direct comparison on how well they identify instrument sounds, focusing on key metrics such as precision, recall and f1-score, especially for the clarinet class, the key aspect of the research goal. The bar chart below (Figure 4.4) visualizes the comparison of the clarinet focused metrics across all the models.



The visual representation supports several import observations. In terms of precision, the CNN model produces a decent result, equalling the precision of the CRNN model, however it struggled with the other metrics, especially recall. This resulted in a lower amount of clarinet sounds being identified. The RNN model obtained a similar recall to the CRNN model and a better result than the CNN. However, it fell short in the precision meaning that it was more sensitive in identifying clarinet sounds than the CNN but was more susceptible to false positives. The CRNN achieved the best balance between the metrics, with both high precision and recall, this achieved the highest F1-score between the models, delivering the most stable performance overall.

Looking at the confusion matrices of each model, a consistent pattern appears across all the models. The miss-classifications of the clarinet were mostly classified as flute or saxophone, not the trumpet. This is due to the distinct temporal and spectral features of the trumpet, such as a bigger frequency spread and brighter harmonics, which when translated into spectrogram-based features, made the trumpet stand out more than the other instruments and easier to identify.

The findings in this study back up the hypothesis that machine learning can identify clarinet sounds against similar sounding instruments. Each model has its trade offs. The CNN is a fast and simple model focusing on spatial features but neglecting temporal data. The RNN focuses on sequential learning which improves recall at the cost of accuracy, The CRNN combines the best of the models but it pays back in increased model complexity and higher training times since CRNN require more computational resources due to its complexity.

Conclusion

This study investigated how different machine learning models – CNN, RNN, and a hybrid CRNN distinguished clarinet sounds from audio files containing instruments with similar characteristics, including the flute, saxophone, and trumpet. The results shows that the hybrid CRNN model performed the best across the models, achieving a more balanced result. This supports the hypothesis that combining both spatial aspects and temporal feature extraction improves clarinet identification.  
  
This was proven by the CRNN’s balanced result. Both CNN and RNN achieved a good result in a certain metric but achieved a poor result in the other metric that make up the F1-score. The CNN achieved decent precision whilst suffering from low recall, and the RNN produced good recall but lacked in precision. The CRNN balanced both metrics which achieved the best F1-score, validating that both spatial and temporal features are necessary for identifying subtle difference between similarly sounding instruments.

The two research questions of this study are addressed by these results. In regard to RQ1, the CRNN was confirmed as the best model for clarinet sound identification, having a definitive advantage over the CNN and CRNN. The different feature extraction techniques discussed in RQ2, the chromas based feature integrated with the CRNN proved to be the most effective. This result shows that chroma characteristics like harmonics and pitch-based representation are more suited to identify the clarinet’s unique acoustic characteristics compared to the feature extraction techniques of the Mel-Spectrogram and MFCC.

Despite the success in answering the research questions, several methodological limitations were identified. To keep with the scope of the study, which is identification of the clarinet against similar instruments, only four instrument classes – clarinet, flute, saxophone, and trumpet were chosen even though the IRMAS dataset is of a polyphonic nature and contains real-world scenarios with overlapping instruments, therefore full advantage of the reliable dataset could not be taken. Another key limitation was reliance on Google Colab for model training. Although Colab contains a good GPU, its strict runtime and resource limitations made it difficult to explore the models’ complexity and size as well as the level of hyperparameter modification. The last limitation is the number and size of models trained. More sophisticated architectures such as Transformers or attention processes may benefit future studies, even if the models used provide a valid baseline comparison. To conclude, this study confirms that machine learning algorithms, particularly hybrid neural network architecture can enhance clarinet sound identification, and offers a solid foundation for future research in clarinet based studies as well as studies revolving around other instruments.