sciendo

Instrument Classification in Musical Audio Signals using Deep Learning

Karlo Borovčak
University of Zagreb, Faculty of Electrical Engineering and Computing

Marina Bagić Babac
University of Zagreb, Faculty of Electrical Engineering and Computing

Abstract

The intersection of artificial intelligence and music technology is creating new possibilities for cultural preservation and innovation. This study aims to utilise this technology by optimising deep learning models for accurate instrument classification, thereby contributing to advancements in music recognition, database organisation, and educational transcription tasks. Using the IRMAS dataset, we evaluated several neural network architectures, including DenseNet121, ResNet-50, and ConvNeXt, trained on log-Mel spectrograms of segmented audio clips to capture the unique acoustic features of each instrument. Results indicate that DenseNet121 achieved the highest classification accuracy, with notable performance in precision, recall, and F1-score compared to other models. However, challenges were observed in recognising instruments with fewer training samples, like the clarinet and cello, underscoring the importance of balanced datasets. While data augmentation techniques only partially addressed class imbalance, the findings offer valuable insights into designing robust music processing systems, highlighting areas for improvement in feature extraction and data handling. This study contributes to the development of AI-driven tools in music, offering potential benefits for cultural and educational growth.

Keywords: deep learning; spectrogram; multi-label classification; instrument recognition

Paper type: Research article

Received: Jun 19, 2024
Accepted: Jan 4, 2025

DOI: 10.2478/crdj-2025-0006

Introduction

The classification of musical instruments from audio signals has garnered significant interest in recent years due to its applications in music information retrieval, automated music transcription, and audio database management. Various studies have employed feature extraction techniques, such as Mel-frequency cepstral coefficients (MFCCs) (Agostini et al., 2003) and log-Mel spectrograms (Umapathy et al., 2007), to transform audio signals into representations suitable for deep learning models. Convolutional Neural Networks (CNNs) have become a popular architecture for this task (Targ et al., 2016), leveraging their ability to learn hierarchical feature representations directly from the spectrograms (Kratimenos et al., 2021). More recently, advanced architectures such as ConvNeXt and DenseNet have been adapted for music processing tasks, demonstrating effectiveness in handling complex audio features and producing reliable classification results (Chakraborty & Parekh, 2018).

These advancements demonstrate the potential of deep learning in musical instrument classes. Class imbalance can skew model performance, favouring well-represented classes over underrepresented ones (Gómez-Cañón et al., 2018), and leading to reduced accuracy for certain instruments (Hernández-Olivan & Beltrán, 2021). Researchers have explored data augmentation techniques (Joder et al., 2009), including pitch shifting and time stretching, as a way to artificially increase data volume and improve model robustness (Rajesh & Nalini, 2020). However, the effectiveness of these methods remains mixed, as they may not fully address the complexity of certain timbral characteristics that distinguish musical instruments in polyphonic audio environments (Park & Lee, 2015).

The purpose of this paper is to evaluate and compare the effectiveness of several deep learning architectures, specifically DenseNet121, ResNet-50, and ConvNeXt, for musical instrument classification, with an emphasis on improving performance for underrepresented instruments. Using the IRMAS dataset, which is well-suited for multi-label classification, we applied log-Mel spectrograms and Mel-frequency cepstral coefficients as input features, implementing data augmentation techniques to address class imbalance (Borotić et al, 2023). Our methodology combines model training on segmented audio clips with careful tuning of neural network hyperparameters to optimise classification performance.

This study contributes to the field by offering a comparative analysis of advanced neural network architectures, underscoring the strengths of DenseNet121 in handling multi-label classification for musical instruments. In traditional single-label classification, each audio segment is classified with only one instrument label **(Chakraborty & Parekh, 2018)**, which limits applicability in polyphonic music where multiple instruments often play simultaneously. Our study employs multi-label classification to address real-world scenarios where multiple instruments may be present in a single audio clip, requiring models to detect all relevant classes accurately.

The remainder of the paper is organised as follows: Section 2 reviews the related work and theoretical background of musical instrument classification; Section 3 details the dataset, preprocessing steps, and data augmentation techniques used. Section 4 explains

the experimental setup, including the architectures and model parameters. Section 5 presents the results of our experiments, while Section 6 discusses these findings, their implications, and potential areas for future research. Finally, Section 7 concludes the paper, summarising our key contributions.

Literature review

Research on musical instrument classification has made significant progress over the last decade, with various deep learning techniques emerging as powerful tools for recognising and differentiating between instruments in audio signals. Early studies focused on feature extraction methods that transform raw audio signals into representations suitable for machine learning models (Khan & Al-Khatib, 2006). For instance, Mel-frequency cepstral coefficients (MFCCs) have been a popular choice due to their ability to capture essential timbral features of audio signals (Mahanta et al., 2021). Similarly, log-Mel spectrograms offer a two-dimensional representation of audio that preserves both spectral and temporal information, making them well-suited for convolutional neural networks (Racharla et al., 2020). Convolutional neural networks (CNNs) have shown promise in learning hierarchical representations from spectrograms, offering notable accuracy improvements over traditional machine learning approaches in instrument classification tasks (Targ et al., 2016).

Recent studies have introduced deeper architectures to capture complex patterns in audio data, such as ResNet and DenseNet, both of which have proven effective in enhancing classification accuracy for audio-related tasks (Chakraborty & Parekh, 2018). ResNet's use of skip connections has enabled deeper networks without the vanishing gradient problem, which previously limited the ability of deep learning models to handle complex data. Similarly, DenseNet's densely connected layers are advantageous for instrument classification, enabling efficient feature propagation throughout the network and resulting in superior feature extraction capabilities (Taenzer et al., 2023). These advancements have expanded the applicability of neural networks in complex audio analysis tasks, including multi-label classification for polyphonic music.

Some researchers have also explored hybrid approaches that combine various feature extraction methods, hoping to leverage unique aspects of each technique to improve classification results. For instance, combining MFCCs with rhythmic patterns or harmonic structures has shown potential in achieving higher accuracy, especially for instruments with complex tonal properties (Hernandez-Olivan & Beltran, 2021). While traditional methods primarily focused on single-label tasks, modern multi-label classification architectures are better suited for real-world scenarios where multiple instruments are often present in a single recording. This shift towards multi-label classification reflects the growing sophistication in the field, driven by improved architectures and data processing methods (Ivezić & Bagić Babac, 2023).

In conclusion, the body of research in musical instrument classification demonstrates a continuous evolution from traditional machine learning approaches to complex deep learning architectures. The introduction of advanced CNN architectures, data

augmentation techniques, and hybrid feature extraction methods has collectively improved model accuracy. However, challenges persist in handling class imbalances and capturing the full complexity of audio signals. This study builds on these developments by applying and comparing state-of-the-art architectures to further refine instrument classification models, ultimately aiming to contribute to the development of more robust music processing systems.

Methodology

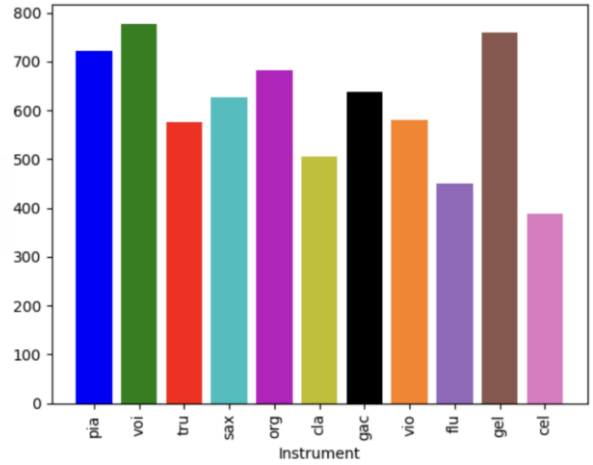
IRMAS dataset

IRMAS (Instrument Recognition in Musical Audio Signals) dataset comprises audio recordings annotated with labels indicating the presence of specific instruments. It is designed for training and testing models for instrument recognition in audio recordings. The considered instruments and their labels are as follows: cello (cel), clarinet (cla), flute (flu), acoustic guitar (gac), electric guitar (gel), organ (org), piano (pia), saxophone (sax), trumpet (tru), violin (vio), and human singing voice (voi) (Bosch et al., 2012). The dataset consists of separate sets for training and validation/testing. The training dataset comprises 6705 audio recordings in 16-bit stereo .wav format sampled at 44.1 kHz, each with a duration of 3 seconds.

The validation data consists of 2,874 files, also in 16-bit stereo. wav format, sampled at 44.1 kHz. Unlike the training data, the validation data is not fixed in length, with durations varying between 5 and 20 seconds. Additionally, a single audio clip in the validation data may contain multiple different instruments. It is worth noting that the training dataset is not perfectly balanced. As illustrated in the accompanying graph (Figure 1), specific instruments such as the electric guitar or human voice have nearly twice as many samples as instruments like the cello or flute.

Figure 1

Distribution of instruments in the IRMAS dataset



Source: Author's illustration

This comprehensive dataset enables the training and evaluation of instrument classification models, offering a diverse range of audio recordings that encompass various musical instruments and performance styles.

Data preprocessing

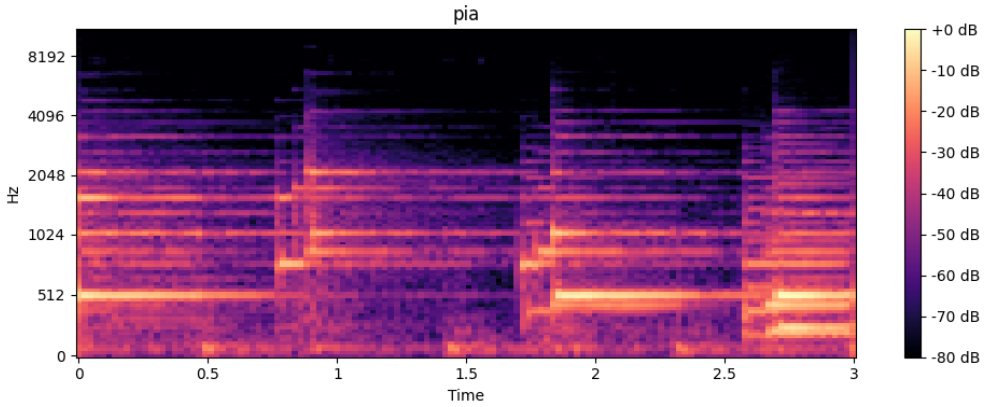
The primary goal of this phase is to cleanse the data of errors, missing values, and unnecessary information, and to transform it into a format suitable for detailed analysis and training the chosen model. This step is essential for ensuring data quality and creating a model that can effectively perform later training on this data.

The first step in preprocessing the dataset involves loading the data. During this phase, we resample audio recordings from the original 44,100 Hz to 22,050 Hz. This resampling is beneficial as it retains most of the information from the recordings while significantly reducing the processing time. For the same reason of efficiency, we also convert stereo sound to mono, as suggested by Rajesh & Nalini (2020). Following the initial adjustments, we segment all files in the training dataset into 1-second clips. This segmentation is also applied to the validation dataset. If a file cannot be evenly divided into 1-second segments, we employ a padding technique, adding zeros to the end of the last segment to reach the desired length. Various segment lengths were experimented with, including 3, 2, 1.5, and 0.5 seconds; however, 1-second segments consistently yielded the best results. A primary focus of our preprocessing is the computation of Mel spectrograms on these segmented audio clips. This method required precise settings for the Short-Time Fourier Transform (STFT). After experimenting with various parameter combinations, we finalised a window size of 1024 and a hop length of 512 and set the number of Mel bands to 128. These parameters were chosen to optimise the balance between computational efficiency and the resolution of the spectrogram, crucial for capturing the distinct features of different musical instruments (Deng et al., 2008).

With these settings, we computed the Mel spectrograms and scaled the results in decibels (dB) to enhance the visibility of the sound spectrum's dynamic range. This step significantly improves our ability to distinguish between different instruments by visualising their unique acoustic signatures more clearly (Chakraborty & Parekh, 2018). The decibel scaling of the spectrograms enables a more detailed and comparative analysis of the audio recordings, which is crucial for accurately classifying the instruments involved.

Figure 2

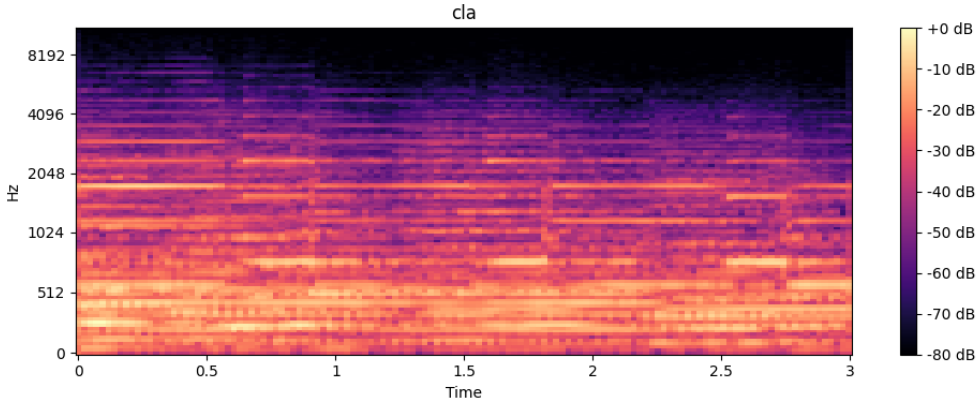
Mel Spectrogram for Piano



Source: Author's illustration

Figure 3

Mel Spectrogram for Clarinet



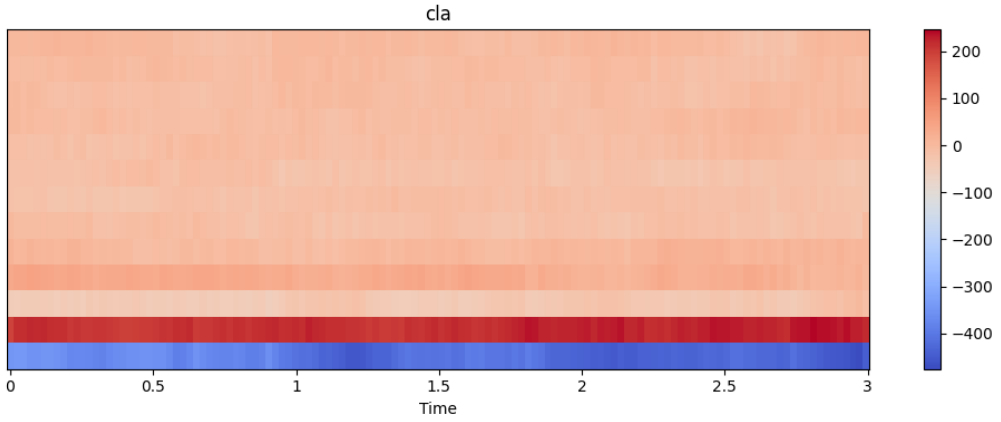
Source: Author's illustration

From Figures 2 and 3, we can see how a Mel spectrogram presents spectral features from the audio signals. Piano is presented with shorter, choppy tones while clarinet is presented with smoother, longer ones. This is a good representation of audio signals that humans perceive, and CNNs can use them to learn effectively.

Following the Mel spectrogram computation, we also extracted Mel-frequency cepstral coefficients (MFCCs) to complement our primary method. This was done by performing an STFT on the audio signals using the same window size and hop length settings. We calculated 13 MFCCs per segment, with the number of Mel bands also set at 128. The choice of these parameters was influenced by the need to provide a high-resolution depiction of the sound while managing computational demands.

Figure 4

MFCCs for Clarinet



Source: Author's illustration

The final step in our preprocessing was the normalisation of both the Mel spectrograms and the MFCCs. This involved scaling the data so that all values fell between 0 and 1, with the minimum value set to 0 and the maximum value set to 1. Normalisation is crucial for reducing bias in the model's learning process due to the range of input feature values, thus ensuring uniform treatment across all features and enhancing the overall performance of our classification model (Profeta & Schuller, 2021). Figure 4 presents the normalised MFCCs for the same clarinet audio signal that was used to calculate the Mel spectrogram. We can see that Mel spectrograms capture more spectral features of audio signals, which is why they are primarily used for music classification tasks.

Data augmentation

Data augmentation is a technique used to artificially enhance the size and quality of training datasets by generating new data points from existing data through various transformations (Kratimenos et al., 2021). This approach addresses common issues in data science related to data scarcity that can impede the robustness and generalisation of models. In our project, after experimenting with different combinations and types of features, we chose to focus on the Mel spectrogram approach. We also decided to implement data augmentation on our training dataset, which provided additional data for training and subsequently improved our model's performance.

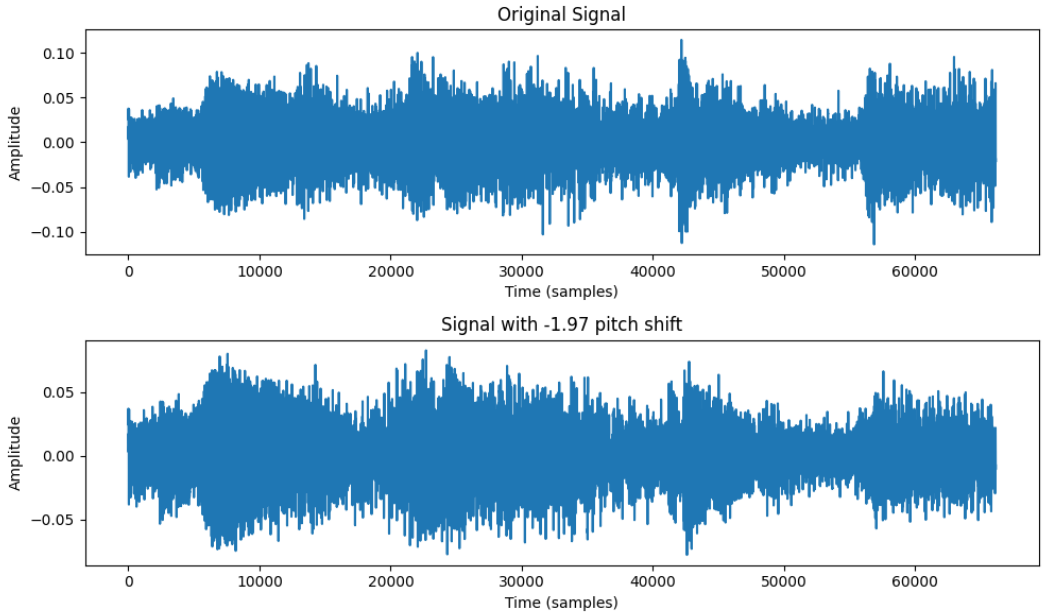
To generate new examples, we experimented with different augmentation values and their combinations to find the most effective outcomes. It was crucial to select augmentation values that neither altered the original audio signal excessively nor were so subtle as to have no impact. The most successful results were obtained by applying time stretching and pitch shifting. For pitch shifting, we generated values from a uniform distribution ranging from -3 to +3. This range enabled us to adjust the pitch of the audio samples upwards or downwards, allowing the model to adapt to different tonal variations and accurately identify the same type of instrument across various pitches, as shown in Figure 5. Time stretching was applied with values ranging from 0.8 to 1.2, altering the

speed of audio clips without affecting their pitch. This ensured that the model could recognise musical instruments regardless of the speed at which they were played.

Interestingly, adding noise to the audio samples did not positively impact the model's performance in our case. This finding led us to focus more on pitch shifting and time stretching, which proved to be more beneficial for enhancing the dataset.

Figure 5

Audio Signal Pitch Shift by -1.97



Source: Author's illustration

By integrating these data augmentation techniques into our preprocessing pipeline, we not only enriched the dataset with realistic variations but also significantly improved the model's ability to perform with higher accuracy when classifying musical instruments from audio signals (Gómez-Cañón et al., 2018). This approach has made the model more robust and applicable across various acoustic environments, thereby enhancing its practical utility in real-world scenarios.

Models

ResNet (Residual Network), introduced in 2015, was a revolutionary architecture that made deep neural networks feasible and effective. The key innovation of ResNet is its use of skip connections, which address the performance degradation problem found in very deep neural networks (Targ et al., 2016). Skip connections in ResNet allow information to bypass certain layers entirely, facilitating both local and global learning, and enabling the training of much deeper networks without a loss in performance. ResNet's capability to train deeper networks effectively made it an optimal choice for tasks such as image classification, object detection, and segmentation.

ConvNeXt is a newer architecture from 2022, building upon and modernising various components of the ResNet architecture (Liu et al., 2022). It started with changing the training method, including optimisation strategies and related hyperparameters. ConvNeXt also shifted from ResNet's basic cells to a layer with a "patchify" strategy, which performs non-overlapping convolutions. Additionally, it increased the network's efficiency to support more parameters. Further enhancements in ConvNeXt include the use of larger kernel sizes (7x7 compared to 3x3), switching from ReLU to GELU as the activation function, and eliminating some normalisation layers. These changes collectively have improved the efficiency and adaptability of ConvNeXt, making it suitable for modern deep learning challenges (Šimić & Bagić Babac, 2024).

DenseNet (Densely Connected Convolutional Network) represents another significant advancement over traditional convolutional network architectures. One of the main advantages of DenseNet is its efficient use of information from all layers of the network. Dense blocks within DenseNet connect each layer to every other layer in a feed-forward fashion (Taenzer et al., 2023). This means that the features extracted at each layer are used as inputs to all subsequent layers, which facilitates the learning of more complex features and improves generalisation. Additionally, DenseNet requires fewer parameters than other architectures, making it less computationally intensive to train and well-suited for adapting to new datasets. DenseNet has been shown to outperform many other architectures in various image classification tasks.

The final model we used for classification and application was the DenseNet121 architecture from the torchvision library, which takes an array log-Mel spectrogram as input with the shape (1, 128, 44). This means the audio is divided into one-second segments, and a normalised log-Mel spectrogram is extracted from each segment with the following parameters: hop length = 512, n_fft = 1024, n_mels = 128. We achieved the final classification by combining three operations. First, we applied the ReLU activation function to the output tensor of the DenseNet121 architecture. Next, we applied an adaptive global pooling layer to reduce the dimensions of the input tensor to size (1, 1). Then, we flattened the resulting values into a one-dimensional tensor. Finally, we employed a fully connected classification network with a sigmoid activation function to obtain the probability for each instrument, thereby performing the final classification.

While training the model, we tuned various hyperparameters, including batch size, learning rate, and the number of epochs, to optimise the training process. The most effective configuration used a batch size of 32 and a learning rate of 0.001. We implemented early stopping based on validation loss to prevent overfitting, with the best results with a patience setting of three epochs. We chose the Adam optimiser for its efficiency and ability to converge quickly. Our loss function of choice was binary cross-entropy (BCE), suitable for the multi-label classification tasks required for our audio files.

Results

We explored various neural network architectures to identify the most effective for classifying musical instruments from audio clips using the IRMAS dataset. We tested DenseNet121, ResNet-50, and ConvNeXt, each trained on normalised log-Mel

spectrograms of one-second audio slices. DenseNet-121 was selected as the final model due to its superior performance, which leverages densely connected layers for complex feature extraction. The DenseNet121 model achieved an F1-score of 0.62, a Hamming Score of 0.89, a Precision of 0.66, and a recall of 0.58. Comparatively, ResNet-50 and ConvNeXt scored slightly lower across these metrics, with ResNet-50 obtaining an F1-score of 0.57 and ConvNeXt 0.61, illustrating their varying capacities for feature processing under the same conditions. All models were validated on the IRMAS validation set, which comprises 2,874 audio files featuring multiple instruments playing in each one. These results underscore the capabilities of DenseNet121 in handling multi-label classification tasks effectively while also highlighting areas for potential improvement in model training and data handling practices.

Table 1

Evaluation Metrics for Each Model

Model	F1-Score	Hamming Score	Precision	Recall
ResNet-50	0.57	0.87	0.63	0.53
ConvNeXt	0.61	0.90	0.68	0.56
DenseNet-121	0.62	0.89	0.66	0.58

Source: Author's results

Further analysis revealed specific challenges in classifying instruments such as the clarinet and cello, possibly due to their lower frequency of occurrence in the validation dataset. This highlights the importance of balanced data representation and may indicate a need for further dataset enrichment, re-evaluation of model hyperparameters, or adjustments to input data processing to enhance classification accuracy for these instruments.

Table 2

Evaluation Metrics for Each Instrument of the DenseNet121 Model

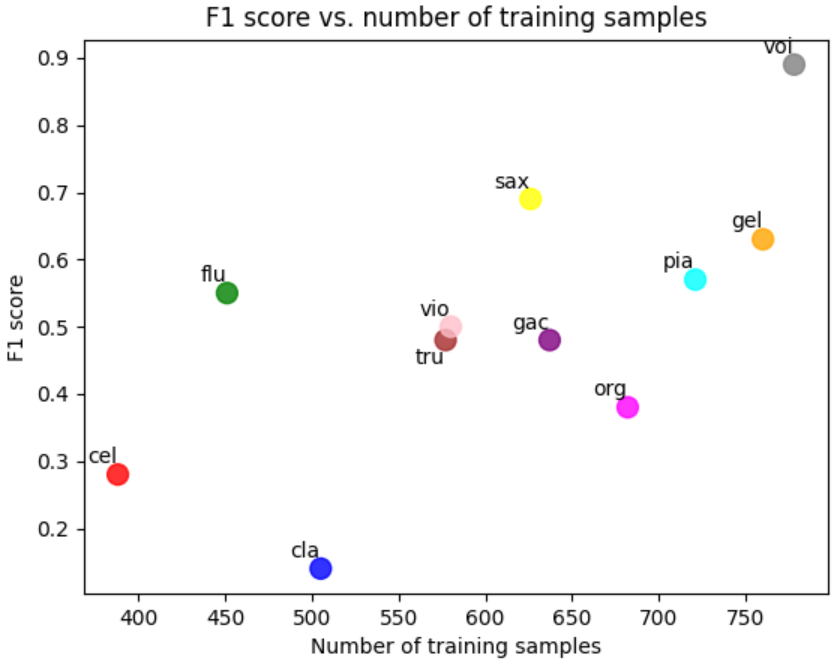
Instrument	F1-Score	Hamming Score	Precision	Recall	Occurrences
Cello	0.28	0.92	0.22	0.41	111
Clarinet	0.14	0.95	0.11	0.19	62
Flute	0.55	0.94	0.48	0.64	163
Acoustic Guitar	0.48	0.86	0.76	0.35	535
Electric Guitar	0.63	0.79	0.74	0.54	942
Organ	0.38	0.88	0.53	0.30	361
Piano	0.57	0.77	0.79	0.45	995
Saxophone	0.69	0.92	0.59	0.84	326
Trumpet	0.48	0.92	0.38	0.64	167
Violin	0.50	0.92	0.46	0.56	211
Voice	0.89	0.92	0.88	0.91	1044

Source: Author's results

As Table 2 suggests, we can observe that some instruments perform significantly better than others. The central hypothesis is that the unbalanced dataset is the cause of these results, and we can test that with a correlation test.

Figure 6

Number of Training Samples vs F1-Score for Each Instrument



Source: Author's illustration

Figure 6 illustrates the relationship between the F1-scores and the number of examples for each instrument in the training dataset. By analysing the Pearson correlation coefficient ($r = 0.6589$, p-value = 0.0274), we confirmed with a significance level of 0.05 that there is an absolute correlation between the F1-score and the number of examples available for each instrument. Following this insight, we attempted to increase the F1-scores of specific underrepresented instruments such as the cello, clarinet, and flute by generating more augmented examples. Unfortunately, these efforts did not yield the expected improvements in F1-scores.

Discussion

Our research focused on enhancing the ability of deep learning models to classify musical instruments from audio samples, utilising the IRMAS dataset. After comparing DenseNet121, ResNet-50, and ConvNeXt, we found that DenseNet121 outperformed the others due to its dense connectivity and efficient feature utilisation, which are crucial for handling complex audio signal data.

We also found a strong correlation ($r = 0.6589$, p-value = 0.0274) between the number of training samples available for each instrument and the F1-score outcomes. This highlights the importance of a balanced dataset in machine learning. Instruments with fewer training samples, like the clarinet and cello, showed lower performance, emphasising the need for balanced representation in the training data.

Efforts to address this imbalance by augmenting data generation for underrepresented instruments were unsuccessful. This suggests that increasing the data quantity may

not be sufficient, and more advanced feature engineering and preprocessing techniques may be needed to capture the complex nature of musical timbres.

Further comparison to related works highlights additional paths for refinement. Studies such as those by Park and Lee (2015) and Hernandez-Olivan and Beltran (2021) demonstrate the utility of combining different feature extraction methods, such as melding Mel spectrograms with rhythmic patterns, to improve instrument classification accuracy. Such hybrid approaches could be investigated to see if they offer a performance benefit in our model setup. Additionally, exploring the features specific to each instrument and their interactions within ensemble settings could offer new insights. For instance, the acoustic properties distinguishing a solo clarinet may differ significantly when it is part of a larger ensemble, and these contextual nuances are crucial for improving classification accuracy in real-world applications (Puh & Bagić Babac, 2023).

While our model shows promise, further work is needed to address dataset imbalances, improve feature extraction methods, and explore innovative architectural tweaks. This continual testing and refinement process is essential for pushing the boundaries of what can be achieved with AI in music technology.

Conclusion

In this paper, we investigated the effectiveness of various deep learning architectures for the task of musical instrument classification in audio signals, explicitly focusing on DenseNet121, ResNet-50, and ConvNeXt. By leveraging the IRMAS dataset and employing log-Mel spectrograms as input features, we assessed each model's precision, recall, and F1-score in a multi-label classification setting. Our findings reveal that DenseNet121 demonstrated superior performance, particularly for well-represented instruments, though challenges persist in accurately classifying underrepresented instruments. These results highlight the importance of both balanced datasets and advanced feature extraction methods in developing reliable instrument classification systems.

Compared to previous research, this study builds upon established CNN architectures and introduces advanced multi-label classification techniques in a deep learning framework. While earlier studies primarily focused on single-label classification tasks or utilised more basic neural networks (Deng et al., 2008), our approach provides a comprehensive evaluation across multiple state-of-the-art architectures. Additionally, data augmentation techniques such as pitch shifting and time stretching were incorporated to address dataset imbalances, although their effectiveness varied, aligning with findings from Gómez-Cañón et al. (2018) and Hernandez-Olivan & Beltran (2021). Our research thus bridges the gap between traditional CNN-based approaches and the potential of multi-label classification in complex audio data.

From a practical perspective, this work has implications for music technology applications, particularly in music information retrieval, music education, and automated transcription systems. Implementing robust multi-label classification systems could enable the development of tools for instrument-specific search

functions in music libraries, facilitate educational platforms with more accurate transcription capabilities, and even assist composers in organising and analysing recordings based on instrumental content. Additionally, insights into handling dataset imbalances may prove valuable for developers working with limited or skewed data.

However, our study has limitations, primarily in addressing the performance gap for underrepresented instruments, where data augmentation alone proved insufficient. Future research could include 'Mamba' models, which integrate wavelet transforms to capture both temporal and spectral details in audio, potentially enhancing the model's ability to distinguish complex timbres. Wavelet decomposition, known for its multi-resolution analysis, may allow better representation of transient and harmonic components of audio signals, making it suitable for intricate timbral analysis. Applying these models could enhance classification accuracy for polyphonic music, particularly when instruments overlap. Further investigation into these techniques may enhance model accuracy and generalizability (Poje et al., 2024), especially for less-represented classes.

References

1. Agostini, G., Longari, M., & Pollastri, E. (2003). Musical Instrument Timbres Classification with Spectral Features. *EURASIP Journal on Advances in Signal Processing*, *2003*(1), 943279. https://doi.org/10.1155/S1110865703210118
2. Borotić, G., Granoša, L., Kovačević, J. & Bagić Babac, M. (2023), Effective Spam Detection with Machine Learning, *Croatian Regional Development Journal,* 3(2), 43-64. https://doi.org/10.2478/crdj-2023-0007
3. Chakraborty, S. S., & Parekh, R. (2018). Improved Musical Instrument Classification Using Cepstral Coefficients and Neural Networks. In J. K. Mandal, S. Mukhopadhyay, P. Dutta, & K. Dasgupta (Eds.), *Methodologies and Application Issues of Contemporary Computing Framework* (pp. 123–138). Springer Singapore. https://doi.org/10.1007/978-981-13-2345-4_10
4. Deng, J. D., Simmermacher, C., & Cranefield, S. (2008). A Study on Feature Analysis for Musical Instrument Classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *38*(2), 429–438. https://doi.org/10.1109/TSMCB.2007.913394
5. Gómez-Cañón, J., Abeßer, J., & Cano, E. (2018, July). *Jazz Solo Instrument Classification with Convolutional Neural Networks, Source Separation, and Transfer Learning*.
6. Gururani, S., Sharma, M., & Lerch, A. (2019). *An Attention Mechanism for Musical Instrument Recognition*. https://doi.org/10.48550/ARXIV.1907.04294
7. Han, Y., Kim, J., & Lee, K. (2017). Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(1), 208–221. https://doi.org/10.1109/TASLP.2016.2632307

8. Hernandez-Olivan, C., & Beltran, J. R. (2021). *Timbre Classification of Musical Instruments with a Deep Learning Multi-Head Attention-Based Model*. https://doi.org/10.48550/ARXIV.2107.06231

9. Joder, C., Essid, S., & Richard, G. (2009). Temporal Integration for Audio Classification With Application to Musical Instrument Classification. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(1), 174–186. https://doi.org/10.1109/TASL.2008.2007613

10. Khan, M. K. S., & Al-Khatib, W. G. (2006). Machine-learning based classification of speech and music. *Multimedia Systems*, *12*(1), 55–67. https://doi.org/10.1007/s00530-006-0034-0

11. Kratimenos, A., Avramidis, K., Garoufis, C., Zlatintsi, A., & Maragos, P. (2021). Augmentation Methods on Monophonic Audio for Instrument Classification in Polyphonic Music. *2020 28th European Signal Processing Conference (EUSIPCO)*, 156–160. https://doi.org/10.23919/Eusipco47968.2020.9287745

12. Ivezić, D. & Bagić Babac, M. (2023). Trends and Challenges of Text-to-Image Generation: Sustainability Perspective, *Croatian Regional Development Journal*, 3(1), 56-77. https://hrcak.srce.hr/file/448285

13. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). *A ConvNet for the 2020s*. https://doi.org/10.48550/ARXIV.2201.03545

14. Mahanta, S. K., Rahman Khilji, A. F. U., & Pakray, P. (2021). Deep Neural Network for Musical Instrument Recognition Using MFCCs. *Computación y Sistemas*, *25*(2). https://doi.org/10.13053/cys-25-2-3946

15. Park, T., & Lee, T. (2015). *Musical instrument sound classification with deep convolutional neural network using feature fusion approach*. https://doi.org/10.48550/ARXIV.1512.07370

16. Poje, K., Brčić, M., Kovač, M., & Bagić Babac, M. (2024), Effect of Private Deliberation: Deception of Large Language Models in Game Play. *Entropy*, 26(6), 524. https://doi.org/10.3390/e26060524

17. Profeta, R., & Schuller, G. (2021). End-to-End Learning for Musical Instruments Classification. *2021 55th Asilomar Conference on Signals, Systems, and Computers*, 1607–1611. https://doi.org/10.1109/IEEECONF53345.2021.9723181

18. Puh, K., & Bagić Babac, M. (2023). Predicting stock market using natural language processing, *American Journal of Business,* 38(2), 41-61. https://www.emerald.com/insight/content/doi/10.1108/AJB-08-2022-0124/full/html,

19. Racharla, K., Kumar, V., Jayant, C. B., Khairkar, A., & Harish, P. (2020). Predominant Musical Instrument Classification based on Spectral Features. *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*, 617–622. https://doi.org/10.1109/SPIN48934.2020.9071125

20. Rajesh, S., & Nalini, N. J. (2020). Musical instrument emotion recognition using deep recurrent neural network. *Procedia Computer Science*, *167*, 16–25. https://doi.org/10.1016/j.procs.2020.03.178

21. Šimić, A. & Bagić Babac, M. (2024). Artificial Intelligence in Classifying and Creating Art: a Survey *International Journal of Student Project Reporting*, 2(1), 59 - 89.

22. Taenzer, M., Mimilakis, S. I., & Abeßer, J. (2023). Deep Learning-Based Music Instrument Recognition: Exploring Learned Feature Representations. In M. Aramaki, K. Hirata, T. Kitahara, R. Kronland-Martinet, & S. Ystad (Eds.), *Music in the AI Era* (Vol. 13770, pp. 32–46). Springer International Publishing. https://doi.org/10.1007/978-3-031-35382-6_4

23. Targ, S., Almeida, D., & Lyman, K. (2016). *Resnet in Resnet: Generalizing Residual Architectures*. https://doi.org/10.48550/ARXIV.1603.08029

24. Umapathy, K., Krishnan, S., & Rao, R. K. (2007). Audio Signal Feature Extraction and Classification Using Local Discriminant Bases. *IEEE Transactions on Audio, Speech and Language Processing*, *15*(4), 1236–1246. https://doi.org/10.1109/TASL.2006.885921

About the authors

Karlo Borovčak received a B.S. in Computer Science from the University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia, where he is pursuing an M.S. in Computer Science. His professional and research interests include software engineering and data science applications. The author can be contacted at karlo.borovcak@fer.hr

Marina Bagić Babac is an Associate Professor at the University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia, where she obtained her Dipl.Ing., M.Sc. and Ph.D. She also obtained an M.Sc. in Journalism from the University of Zagreb's Faculty of Political Science. She is actively involved in several international projects in the field of artificial intelligence. She serves as a program committee member for several international scientific conferences and journals and as a reviewer for numerous international journals. Her research interests include artificial intelligence, machine learning, natural language processing, and social network analysis. The author can be contacted at marina.bagic@fer.hr