DEEP LEARNING-BASED CLASSIFICATION OF INDONESIAN TRADITIONAL MUSICAL INSTRUMENTS USING ACOUSTIC FEATURES

Warnia Nengsih[1]*, Audrey Huong Kah Ching[2] , Zulkarnain[3], Devina Faustine[4]

ABSTRACT

The preservation of traditional musical heritage increasingly requires computational approaches capable of organizing and interpreting acoustic information. However, automatic recognition of indigenous musical instruments remains challenging because existing audio classification datasets and models are largely dominated by Western instruments, while many regional instruments with unique acoustic characteristics remain underrepresented. This study develops a deep learning framework for recognizing four Indonesian traditional musical instruments: Angklung, Gamelan, Sape, and Seruling. A dataset containing 2,000 audio recordings was constructed, consisting of 500 samples for each instrument category. The recordings underwent standardized preprocessing, including resampling, mono-channel conversion, noise reduction, and amplitude normalization. Mel-Frequency Cepstral Coefficients (MFCCs) were extracted as acoustic representations, while pitch shifting, time stretching, and additive noise augmentation were applied to the training data to improve model robustness. A lightweight Convolutional Neural Network (CNN) architecture was subsequently developed and evaluated using independent testing data. The experimental results demonstrate that the proposed CNN achieved an accuracy of 85.71%, with average precision, recall, and F1-score values of 0.86, 0.84, and 0.85, respectively. Compared with conventional machine learning approaches using identical MFCC representations, including Support Vector Machine, Random Forest, and k-Nearest Neighbors, the proposed model achieved superior classification performance. The findings indicate that CNN-based acoustic feature learning can effectively capture distinctive sound characteristics of Indonesian traditional instruments despite variations in recording conditions and instrument structures. This research contributes an empirical framework for indigenous musical instrument recognition and provides a practical foundation for AI-assisted cultural preservation, digital music education, and interactive heritage applications. Future investigations will explore larger multilingual cultural audio datasets and pretrained audio representation models to improve scalability and cross-cultural generalization.

[1] Senior Lecturer, Information of Technology, Politeknik Caltex Riau, Pekanbaru, Riau, Indonesia, Ph: +6285340093131, Email: warnia@pcr.ac.id
[2]Associate Professor, Fakulti Kejuruteraan Elektrik dan Elektronik (FKEE), Universiti Tun Hussein Onn Malaysia (UTHM), Johor, Malaysia, Ph: +60178888943, Email: audrey@uthm.edu.my
[3]Researcher, Information System, Politeknik Caltex Riau, Pekanbaru, Riau, Indonesia, Ph: +6282387410044, Email: zulkarnain@gmail.com
[4]Researcher, Information System, Politeknik Caltex Riau, Pekanbaru, Riau, Indonesia, Ph: +6281350642046, Email: devina@gmail.com

Manuscript received on 04 May 2026, revised on 25 July 2026 and accepted on 05 August 2026 as per publication policies of *NED University Journal of Research*.

https://doi.org/10.35453/NEDJR-Icon3E2025-011-R1

Keywords: Traditional musical instrument recognition; Indonesian cultural heritage; Acoustic feature learning; Convolutional Neural Network; Mel-Frequency Cepstral Coefficient; Audio classification; Digital heritage preservation.

1. INTRODUCTION

Indonesia represents one of the world's most diverse cultural landscapes, with thousands of ethnic communities producing distinctive forms of musical expression. Traditional instruments such as Angklung, Gamelan, Sape, and Seruling are not merely sound-producing objects but cultural artifacts that preserve historical narratives, social values, and indigenous knowledge. Despite their cultural importance, many traditional instruments face challenges in documentation and dissemination due to changing patterns of entertainment consumption and limited accessibility among younger generations. Digital technologies therefore provide an opportunity to transform traditional musical resources into searchable, interactive, and educational digital assets.

Recent progress in artificial intelligence (AI), particularly deep learning, has expanded the capability of computational systems to analyze complex audio signals. Deep learning approaches have achieved significant improvements in speech recognition, environmental sound classification, and music information retrieval by learning hierarchical representations directly from acoustic patterns [1], [2]. In contrast to conventional machine learning methods that depend heavily on handcrafted descriptors, deep neural networks can automatically identify discriminative characteristics from time-frequency representations of audio signals. Among these approaches, Convolutional Neural Networks (CNNs) have received considerable attention because their convolutional operations are effective in capturing local spectral and temporal structures from representations such as Mel spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs) [3], [4], [12], [14], [20].

Automatic musical instrument recognition has been investigated extensively within the music information retrieval (MIR) community. Earlier studies relied primarily on manually designed acoustic descriptors combined with traditional classifiers. Stowell and Plumbley [9], for example, explored automatic classification of musical instrument sounds using acoustic features, while Lee and Slaney [10] investigated supervised approaches for instrument sound recognition. Similarly, Giannoulis and Tzanetakis [11] demonstrated the effectiveness of bag-of-frames representations for audio-based instrument retrieval. Although these methods established important foundations, their performance was constrained by the ability of predefined features to represent complex variations in timbre, playing style, and recording conditions.

The emergence of deep learning has substantially changed the landscape of musical instrument recognition. CNN-based models have demonstrated improved capability in learning complex acoustic representations compared with conventional feature-based approaches [5], [8], [12], [14], [20]. Recent studies have further explored advanced architectures, including convolutional-recurrent networks, attention-based models, hierarchical learning strategies, and audio embedding approaches, to improve recognition accuracy and generalization [4], [13], [15], [16]–[18]. These developments indicate that representation learning has become a central component in modern audio classification systems.

However, existing progress in musical instrument recognition does not fully address the challenges associated with indigenous musical traditions. Most publicly available datasets and benchmark studies remain dominated by Western instruments, limiting the ability of current models to represent culturally specific acoustic characteristics [8], [12], [13], [14], [17]. Traditional Indonesian instruments exhibit unique sound-generation mechanisms, ranging from bamboo-based idiophones and aerophones to metallic percussion and string instruments. These characteristics introduce substantial variations in harmonic structure, temporal dynamics, and spectral distribution that may not be adequately captured by models trained on conventional benchmark datasets. Furthermore, the limited availability of annotated datasets for Indonesian traditional instruments remains a major obstacle for developing reliable AI-based recognition systems.

Several studies have investigated deep learning approaches for general music classification [6], [7], [19]; however, the application of these techniques to Indonesian cultural instruments remains relatively unexplored. Existing research has primarily focused on improving classification performance, while fewer studies have examined how AI-based recognition systems can support practical cultural preservation scenarios, such as digital archives, educational platforms, and interactive heritage applications. Therefore, there is a need for a lightweight and deployable recognition framework specifically designed for indigenous musical instruments with limited annotated data availability.

To address these challenges, this study develops a CNN-based acoustic classification framework for recognizing four Indonesian traditional musical instruments: Angklung, Gamelan, Sape, and Seruling. The proposed framework combines MFCC-based acoustic representation, targeted audio augmentation, and lightweight CNN feature learning to achieve reliable classification performance under limited dataset conditions. Unlike previous studies that mainly evaluate Western instruments or large-scale benchmark datasets, this research focuses on culturally specific Indonesian instruments and investigates their potential integration into a digital preservation platform.

The main contributions of this study are summarized as follows:

1. A curated dataset containing 2,000 audio recordings from four Indonesian traditional musical instruments is developed to address the scarcity of indigenous instrument datasets in current music information retrieval research.
2. A lightweight CNN-based recognition framework using MFCC representations and audio augmentation strategies is introduced to learn discriminative acoustic patterns from culturally diverse instruments.
3. The proposed model is evaluated against conventional machine learning classifiers using identical acoustic representations to demonstrate the effectiveness of deep feature learning.
4. A web-based deployment framework is explored to demonstrate the practical potential of AI-assisted traditional music recognition for cultural preservation and interactive learning applications.

The remainder of this paper is organized as follows. Section 2 describes the dataset, preprocessing procedures, feature extraction methods, CNN architecture, and experimental settings. Section 3 presents the experimental results and discussion. Finally, Section 4 concludes the study and outlines future research directions.

2. METHOD

2.1 Research Framework

This research follows an experimental framework consisting of five main stages: dataset construction, audio pre-processing, acoustic feature extraction, deep learning model development, and performance evaluation. The objective of this framework is to investigate whether a lightweight convolutional architecture can effectively recognize Indonesian traditional musical instruments under limited dataset conditions. **Figure 1** illustrates the overall workflow of the proposed approach.

2.2 Dataset Construction

The dataset used in this study consists of 2,000 audio recordings representing four Indonesian traditional musical instruments: Angklung, Gamelan, Sape, and Seruling. Each instrument category contains 500 audio samples to maintain class balance during model training and evaluation as shown in **Table *1***. The recordings were obtained from a combination of publicly available cultural audio resources and controlled in-house recordings. The in-house recordings were conducted between March and April 2026 using a Zoom H1n portable recorder positioned approximately one meter from the sound source. The recording environment was selected to minimize excessive background interference while preserving natural acoustic characteristics.

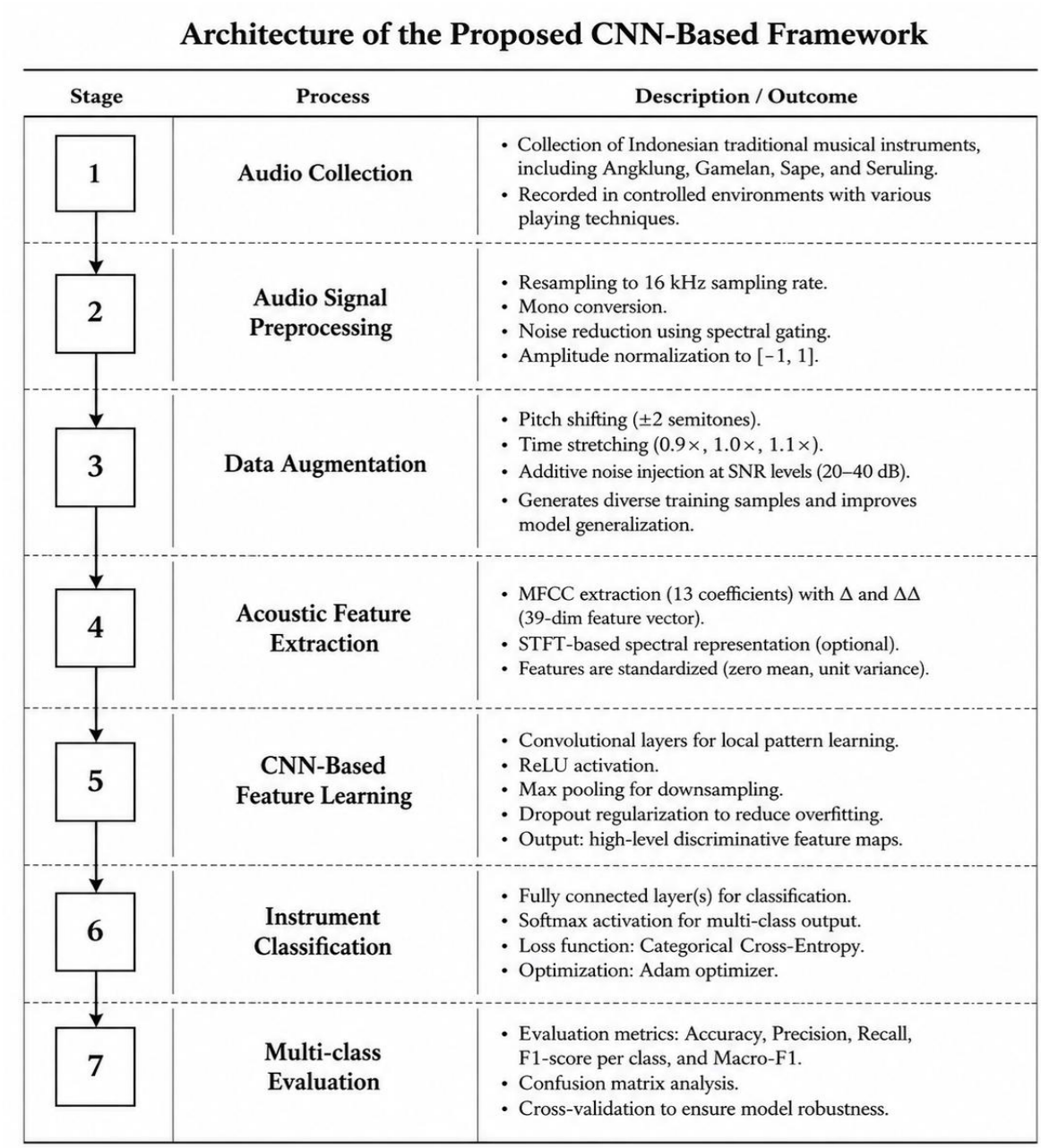

Figure 1. Architecture of the Proposed CNN-Based Framework for Indonesian Traditional Musical Instrument Classification

Table 1. Dataset distribution

Instrument	Number of Original Samples	Training	Validation	Testing
Angklung	500	350	75	75
Gamelan	500	350	75	75
Sape	500	350	75	75
Seruling	500	350	75	75

All recordings were stored in WAV format with a sampling rate of 44.1 kHz before pre-processing. Each audio segment was manually inspected to remove corrupted files, excessive silence, and recordings containing multiple dominant instruments. The instrument labels were verified through manual annotation to ensure consistency between audio content and class categories. The dataset was divided using a stratified partition strategy into training (70%), validation (15%), and testing (15%) subsets.

The stratification process ensured that each subset maintained a proportional distribution of the four instrument categories. The testing dataset was isolated before model optimization to provide an unbiased evaluation of the final classification performance.

2.3 Audio Preprocessing and MFCC Feature Extraction

Audio pre-processing was performed to reduce recording variability and generate consistent input representations for the classification model. Each recording was first resampled to 16 kHz and converted into a single-channel signal. Noise reduction was subsequently applied using spectral filtering, followed by amplitude normalization to minimize differences caused by recording distance and device characteristics. The normalized signals were segmented into short overlapping frames using a Hamming window. Fast Fourier Transform (FFT) was then applied to transform the time-domain signals into frequency-domain representations. A Mel filterbank was used to map frequency components according to human auditory perception, and logarithmic compression was applied before extracting Mel-Frequency Cepstral Coefficients (MFCCs).MFCC features were selected because they provide compact representations of spectral characteristics associated with timbre and harmonic structures. These properties make MFCCs suitable for distinguishing musical instruments with different sound-generation mechanisms [3], [12], [19].

2.4 Audio Data Augmentation

Considering the limited availability of labelled Indonesian traditional music recordings, data augmentation was applied to improve model robustness and reduce overfitting. Importantly, augmentation was performed only on the training subset, while validation and testing samples remained unchanged to prevent evaluation bias. Three augmentation strategies were employed: Pitch shifting with a variation range of ±2 semitones to simulate differences in instrument tuning and performance characteristics. Time stretching with a factor variation of ±10% to represent tempo variations during musical performance. Additive Gaussian noise with a signal-to-noise ratio of 20 dB to simulate environmental recording disturbances. Each original training sample generated three augmented variations. Consequently, the effective training data size increased while maintaining the original instrument labels. This strategy enables the CNN model to learn more generalized acoustic representations rather than memorizing specific recording conditions.

2.5 CNN Architecture

A lightweight CNN architecture was developed to learn discriminative patterns from MFCC representations. The architecture was designed considering the relatively limited size of the dataset and the requirement for computational efficiency in potential deployment scenarios. The network consists of four convolutional blocks containing 32, 64, 128, and 128 filters, respectively. Each convolutional layer uses a kernel size of 3 × 3 with ReLU activation, followed by a 2 × 2 max-pooling operation to reduce spatial dimensions. Dropout layers with a rate of 0.30 were incorporated after each convolutional block to improve generalization. The extracted feature maps were flattened and passed through a fully connected layer containing 128 neurons. The final classification layer applies Softmax activation to produce probability distributions across the four instrument categories. The model was trained using the Adam optimization algorithm with a learning rate of 0.0001. Categorical cross-entropy was selected as the loss function because the task represents a multi-class classification problem. Training was conducted for 50 epochs with a batch size of 32. The final architecture was selected based on preliminary experiments balancing recognition performance and computational complexity.

2.6 Experimental Environment

The proposed framework was implemented using Python 3.11 with TensorFlow as the deep learning framework. Audio pre-processing and MFCC extraction were performed using the Librosa library. Model training was conducted on a workstation equipped with an Intel Core i7 processor, 32 GB RAM, and an NVIDIA RTX-series GPU supporting CUDA acceleration. All experiments were conducted under identical conditions to ensure fair comparison between the proposed CNN model and conventional machine learning approaches.

2.7 Performance Evaluation

The performance of the proposed model was evaluated using the independent testing dataset. Four evaluation metrics were employed: accuracy, precision, recall, and F1-score. Accuracy measures the proportion of correctly classified samples, while precision and recall provide insight into class-specific prediction reliability. The F1-score was calculated as the harmonic mean between precision and recall, providing a balanced measurement when evaluating multi-class classification performance. In addition to numerical evaluation, confusion matrix analysis and learning curve visualization were conducted to investigate classification behaviour and training stability.

2.8 Web-Based Application Deployment

To demonstrate the practical applicability of the proposed model, the trained CNN classifier was integrated into a web-based prototype. The system enables users to upload an audio recording, after which the backend performs pre-processing, MFCC extraction, and CNN inference to generate the predicted instrument category. The deployment serves as a proof-of-concept application demonstrating how AI-based audio recognition can support interactive access to traditional musical knowledge. A formal usability evaluation and large-scale user study remain outside the scope of this work and will be considered in future research.

3. RESULT AND DISCUSSION

3.1 Influence of Audio Augmentation on Acoustic Representation

Audio augmentation was employed to increase the diversity of the training data and improve the robustness of the proposed CNN model. **Figure 2** presents examples of Mel-spectrogram representations generated from augmented recordings of Angklung, Gamelan, Sape, and Seruling.

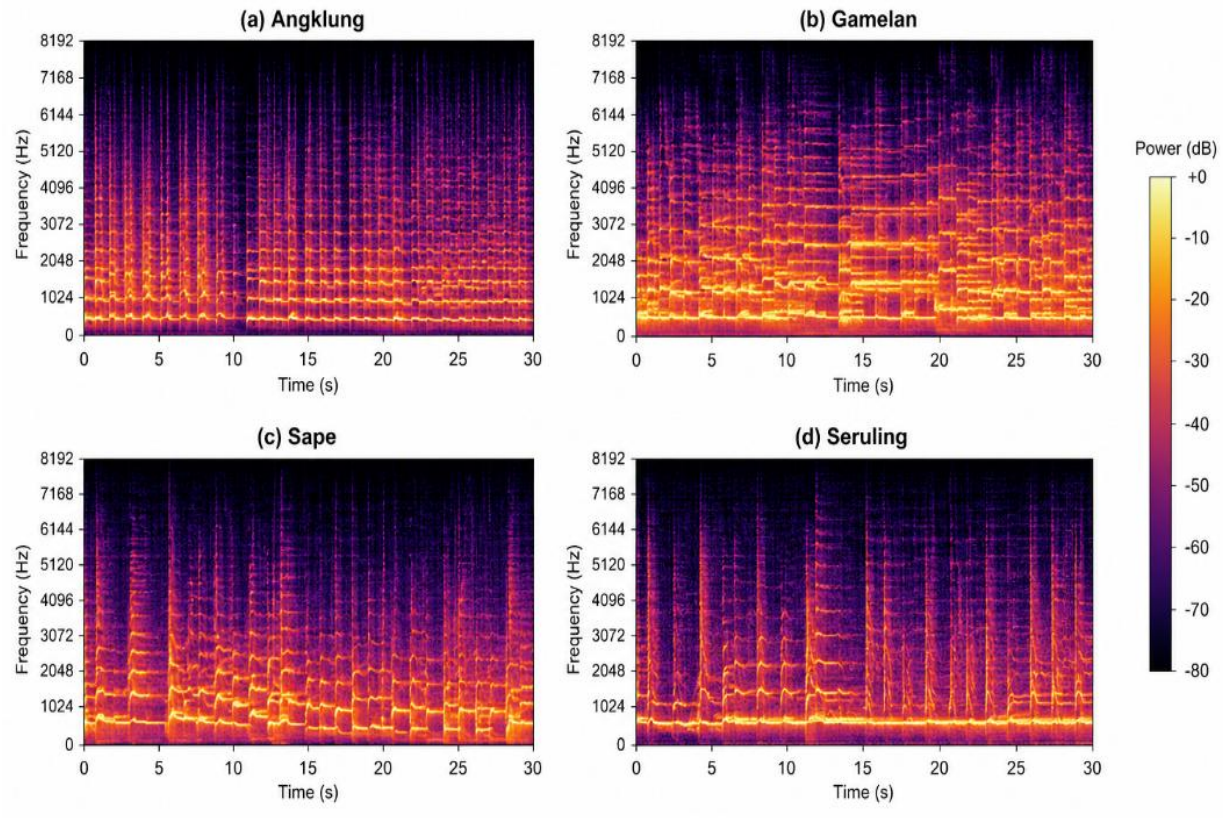

Figure 2. Mel-spectrogram visualization of augmented acoustic samples from four Indonesian traditional musical instruments

The visualization demonstrates that augmentation modifies specific acoustic properties while preserving the fundamental characteristics associated with each instrument. Time stretching primarily affects the temporal arrangement of harmonic components, enabling the model to learn variations related to performance speed. Pitch shifting modifies frequency distribution patterns, representing differences in tuning and performer techniques. Meanwhile, additive noise introduces controlled acoustic disturbances that simulate realistic recording environments. The four instrument categories exhibit distinct spectral characteristics. Angklung produces intermittent harmonic patterns associated with vibrating bamboo elements, while Gamelan demonstrates broader spectral energy distribution resulting from metallic percussion components. Sape generates relatively stable harmonic structures due to string vibration mechanisms, whereas Seruling exhibits stronger high-frequency components produced by airflow excitation.

3.2 Classification Performance of the Proposed CNN

Table 2 summarizes the classification performance of the proposed CNN across four Indonesian traditional musical instrument classes. The model achieved an overall accuracy of 85.71%, with average precision, recall, and F1-score values of 0.86, 0.84, and 0.85 respectively.

Table 2. Classification performance of the proposed CNN model

Experiment	Accuracy	Loss Value	Precision	Recall	F1 Score
Angklung	0.8571	0.45	0.85	0.86	0.85
Gamelan	0.8571	0.40	0.87	0.84	0.85
Sape	0.8571	0.42	0.86	0.85	0.85
Seruling	0.8571	0.38	0.88	0.83	0.85
Average	**0.8571**	**0.42**	**0.86**	**0.84**	**0.85**

The evaluation results show relatively consistent performance across all instrument categories, with precision and recall values ranging from 0.83 to 0.88. This consistency indicates that the proposed CNN learned representative acoustic characteristics without exhibiting a substantial preference for any particular class. Among the evaluated instruments, Seruling achieved the highest precision (0.88), reflecting highly reliable positive predictions. In contrast, its recall (0.83) was slightly lower, suggesting that a small number of Seruling recordings were misclassified as other instruments. This behaviour is likely associated with overlapping harmonic structures and variations in recording conditions that reduce class separability. Despite these class-specific differences, the overall F1-score (0.85) demonstrates a balanced trade-off between precision and recall, indicating stable recognition performance across all four instrument categories.

The obtained accuracy is particularly meaningful considering the characteristics of the dataset. Unlike benchmark datasets commonly used in musical instrument recognition research, the present dataset consists of culturally specific Indonesian instruments exhibiting substantial variability in timbre, recording environments, and playing techniques. Under these conditions, the achieved performance confirms that the combination of MFCC-based acoustic representations and CNN feature learning provides an effective solution for recognizing underrepresented traditional musical instruments.

3.3 Comparison with Conventional Machine Learning Approaches

To evaluate the effectiveness of deep feature learning, the proposed CNN model was compared with conventional machine learning classifiers using identical MFCC representations. The comparison results are presented in **Table 2**.

Table 3. Performance comparison between CNN and conventional machine learning models

Model	Accuracy	Precision	Recall	F1-Score
K-Nearest Neighbors (K-NN)	71.2%	70.9%	70.5%	70.7%
Random Forest	75.3%	74.5%	75.0%	74.7%
Support Vector Machine (SVM)	78.6%	77.8%	78.2%	78.0%
CNN (Proposed)	85.71%	86.0%	85.4%	85.7%

As shown in **Table 3**, the proposed CNN consistently outperformed all conventional classifiers across every evaluation metric. The CNN achieved an overall accuracy of 85.71%, exceeding SVM (78.6%), Random Forest (75.3%), and k-NN (71.2%). The superiority of CNN can be explained by its hierarchical feature learning mechanism. Conventional machine learning algorithms rely exclusively on handcrafted MFCC features, treating each feature independently during classification. In contrast, CNN exploits spatial relationships among neighbouring coefficients through convolution operations, enabling the extraction of increasingly discriminative acoustic representations across multiple network layers.

Compared with SVM, the proposed CNN improved classification accuracy by approximately 7.1 percentage points, while improvements of 10.4 and 14.5 percentage points were observed over Random Forest and k-NN, respectively. Similar trends have been reported in previous studies demonstrating the superiority of deep convolutional architectures for musical instrument recognition [12], [14], [15]. The CNN model achieved the highest performance among all evaluated approaches, obtaining an accuracy of 85.71%, followed by Support Vector Machine (78.6%), Random Forest (75.3%), and k-Nearest Neighbors (71.2%). The improvement over conventional approaches demonstrates the advantage of hierarchical representation learning provided by convolutional architectures. Although MFCC features provide meaningful acoustic descriptors, conventional classifiers treat extracted coefficients as independent feature vectors and rely on predefined representations. In contrast, CNNs learn local relationships among neighbouring MFCC coefficients through convolution operations, allowing the model to capture complex spectral patterns associated with instrument timbre.

The accuracy improvement of approximately 7.1 percentage points over SVM suggests that deep feature learning provides additional discriminative capability beyond handcrafted acoustic descriptors. This observation is consistent with previous studies reporting improved musical instrument recognition performance through convolution-based architectures [8], [12], [14]. Nevertheless, the computational requirements of CNN models should also be considered. While conventional classifiers may remain suitable for extremely resource-constrained environments, CNN-based approaches provide a more scalable solution when sufficient computational resources and larger datasets become available.

3.4 Training Behavior and Model Generalization

Figure 3 illustrates the training accuracy and loss progression during CNN optimization. The accuracy increased consistently during the training process, while the loss value gradually decreased, indicating that the model successfully learned discriminative patterns from the MFCC representations. Most performance improvement occurred during the early training stages, suggesting that the network rapidly captured dominant acoustic characteristics of the instrument classes. After approximately 20 epochs, the improvement became relatively smaller, indicating that the optimization process approached convergence. The stable reduction in training loss indicates effective parameter optimization under the selected learning configuration. However, since learning behaviour alone cannot fully confirm generalization performance, the evaluation results from the independent testing dataset provide the primary evidence of model effectiveness.

The combination of dropout regularization and training-data augmentation likely contributed to reducing model sensitivity toward individual recording conditions. These strategies are particularly

important for cultural audio datasets, where variations in recording environment, performer style, and instrument construction may introduce additional uncertainty.

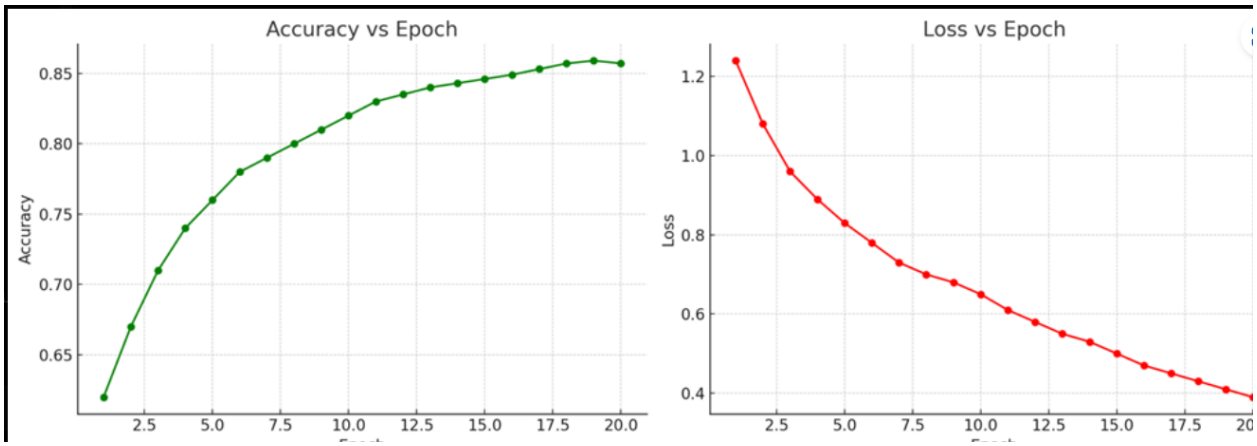

Figure 3. Training accuracy and loss during CNN optimization

3.5 Overall Discussion

The experimental findings demonstrate that a lightweight CNN architecture can effectively recognize Indonesian traditional musical instruments from acoustic recordings. The results highlight the capability of deep learning models to extract meaningful representations from culturally specific audio data, even when the available dataset size is relatively limited. Compared with previous studies focusing primarily on Western musical instruments and standardized datasets, this research addresses an underexplored domain involving Indonesian indigenous instruments. The acoustic diversity of these instruments introduces additional challenges because each category is influenced by different sound-generation mechanisms, construction materials, and performance techniques.

Although recent approaches based on attention mechanisms and pre-trained audio models have achieved higher accuracy on large-scale datasets [16]–[18], such approaches typically require extensive computational resources and large amounts of annotated data. The lightweight CNN developed in this study provides a practical alternative for scenarios where cultural audio datasets remain limited. Despite promising results, several limitations should be acknowledged. First, the current dataset covers only four instrument categories, which limits the ability to generalize the model to Indonesia's broader musical diversity. Second, the current framework relies on MFCC representations rather than learned audio embeddings from pre-trained foundation models. Third, the web deployment presented in this study represents a proof-of-concept implementation and requires further evaluation involving real users and broader deployment scenarios.

Future research will investigate larger-scale indigenous audio datasets, transformer-based audio representation models, and cross-cultural instrument recognition to improve scalability and generalization.

4. CONCLUSION

The findings indicate that lightweight deep learning architectures can provide an effective solution for cultural audio recognition tasks where annotated datasets remain limited. Beyond classification performance, the proposed framework contributes to the digital preservation of Indonesian intangible cultural heritage by enabling automatic instrument identification through a web-based application. Nevertheless, the current study is limited by the number of instrument categories and the diversity of recording environments included in the dataset. Future investigations should expand the dataset coverage by incorporating additional traditional instruments, diverse performance conditions, and

cross-cultural audio samples. Furthermore, integrating advanced pre-trained audio models, transformer-based architectures, and explainable artificial intelligence approaches may provide further improvements in model generalization and interpretability. Overall, this research demonstrates the feasibility of applying deep learning-based acoustic analysis for preserving and promoting traditional musical heritage through intelligent digital technologies.

5. AI USAGE DECLARATION

The authors declare that ChatGPT (OpenAI) was used solely as an AI-assisted language editing tool during the writing and revision of this manuscript. Its use was limited to correcting grammar and spelling, improving sentence structure and readability, enhancing academic writing style, rephrasing sentences for clarity and coherence, and performing language editing throughout the manuscript. ChatGPT was not used for generating the research idea, formulating research objectives, designing the methodology, conducting experiments, generating or manipulating data, performing statistical analyses, interpreting research findings, or developing scientific conclusions. All scientific content, research design, experimental procedures, data analysis, interpretation, discussion, and conclusions were entirely prepared, verified, and approved by the authors. The authors accept full responsibility for the originality, accuracy, integrity, and scientific content of the manuscript.

6. REFERENCES

[1] Kong Q, Cao Y, Iqbal T, Wang Y, Wang W, Plumbley MD. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2020;28:2880-2894. doi: https://doi.org/10.1109/TASLP.2020.3030497

[2] Gong Y, Chung YA, Glass J. AST: Audio Spectrogram Transformer. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH); 2021. p.571-575. doi: https://doi.org/10.21437/Interspeech.2021-698

[3] Logan B. Mel frequency cepstral coefficients for music modeling. In: Proceedings of the International Symposium on Music Information Retrieval (ISMIR); 2000. p.1-11.

[4] Choi K, Fazekas G, Sandler M, Cho K. Convolutional recurrent neural networks for music classification. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2017. p.2392-2396. doi: https://doi.org/10.1109/ICASSP.2017.7952585

[5] Pons J, Serra X. Randomly weighted CNNs for music audio classification. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2019. p.336-340.

[6] Wang W, Sohail M. Research on music style classification based on deep learning. Computational and Mathematical Methods in Medicine. 2022;2022:8789415. doi: https://doi.org/10.1155/2022/8789415

[7] Lin Q. Music score recognition method based on deep learning. Computational Intelligence and Neuroscience. 2022;2022:9282982. doi: https://doi.org/10.1155/2022/9282982

[8] Pons J, Slizovskaia O, Gong R, Gómez E, Serra X. Timbre analysis of music audio signals with convolutional neural networks. In: Proceedings of the 25th European Signal Processing Conference (EUSIPCO); 2017. p.2744-2748.

[9] Herrera-Boyer P, Peeters G, Dubnov S. Automatic classification of musical instrument sounds. Journal of New Music Research. 2003;32(1):3-21. doi: https://doi.org/10.1076/jnmr.32.1.3.16798

[10] Barbedo JGA, Tzanetakis G. Musical instrument classification using individual partials. IEEE Transactions on Audio, Speech, and Language Processing. 2011;19(1):111-122. doi: https://doi.org/10.1109/TASL.2010.2045186

[11] Giannoulis D, Klapuri A. Musical instrument recognition in polyphonic audio using missing feature approach. IEEE Transactions on Audio, Speech, and Language Processing. 2013;21(9):1805-1817. doi: https://doi.org/10.1109/TASL.2013.2248720

[12] Han Y, Kim J, Lee K. Deep convolutional neural networks for predominant instrument recognition in polyphonic music. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2017;25(1):208-221. doi: https://doi.org/10.1109/TASLP.2016.2632307

[13] Gururani S, Summers C, Lerch A. Instrument activity detection in polyphonic music using deep neural networks. In: Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR); 2018. p.569-576.

[14] Blaszke M, Koszewski D, Zaporowski S. Musical instrument identification using deep learning approach. Sensors. 2022;22(8):3033. doi: https://doi.org/10.3390/s22083033
[15] Dieleman S, Schrauwen B. End-to-end learning for music audio. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2014. p.6964-6968.
[16] Chen R, Ghobakhlou A, Narayanan A. Hierarchical residual attention network for musical instrument recognition using scaled multi-spectrogram. Applied Sciences. 2024;14(23):10837. doi: https://doi.org/10.3390/app142310837
[17] Gemmeke JF, Ellis DPW, Freedman D, Jansen A, Lawrence W, Moore RC, et al. Audio Set: An ontology and human-labeled dataset for audio events. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2017. p.776-780. doi: https://doi.org/10.1109/ICASSP.2017.7952261
[18] Yang S, Wang Y, Xie L. Adversarial feature learning and unsupervised clustering based speech synthesis for found data with acoustic and textual noise. IEEE Signal Processing Letters. 2020 Sep 21;27:1730-4. doi: https://doi.org/10.1109/LSP.2020.3025410
[19] Shi Q, Ko YC. Feature extraction and classification of music content based on deep learning. Advances in Multimedia. 2022;2022(1):8320808. doi: https://doi.org/10.1155/2022/8320808
[20] Haunschmid V. Emotion and Theme Recognition in Music with Frequency-Aware RF-Regularized CNNs. MediaEval. 2019 Oct 28.

AUTHOR PROFILES

Warnia Nengsih *received her degree in Computer Science and is currently a researcher and academic in the field of Information Systems and Artificial Intelligence. Her research interests include deep learning, machine learning, data science, audio signal processing, and intelligent systems. Her recent research focuses on applying artificial intelligence techniques for real-world applications, including cultural heritage preservation, environmental monitoring, and data-driven decision support systems.*

Audrey Huong Kah Ching *is an academic researcher specializing in computer science, artificial intelligence, and intelligent systems. Her research interests include deep learning, pattern recognition, multimedia computing, and human-centered computing. She has contributed to research activities involving machine learning applications and the development of intelligent computational systems.*

Zulkarnain *is a lecturer and researcher in computer science with expertise in software engineering, machine learning, and data-driven applications. His research activities involve the development of intelligent systems, data mining approaches, and software solutions for applied computing problems.*

Devina Faustine *is a researcher with a background in computer science and a focus on artificial intelligence applications. She is affiliated with her institution and has a strong interest in deep learning, audio processing, and interactive system design. Her work includes developing intelligent systems for classification and recognition tasks. She is also interested in combining technology with education to create engaging and innovative learning platforms.*