Vol. 11, No. 1, April 2026, Page. 288-297
https://journal.uin-alauddin.ac.id/index.php/instek
p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

Comparative Evaluation of LSTM Variants for Traditional Musical Instrument Audio Classification Using MFCC Features

Ridwan Andi Kambau[*1], Muhammad Syawal Idil Fitrah Baharuddin[2]

[1]Informatics Engineering, Faculty of Science and Technology, UIN Alauddin Makassar, Indonesia
Informatics Engineering, Faculty of Engineering, Universitas Hasanuddin
Email: [1]ridwan.kambau@uin-alauddin.ac.id, [2]baharuddinmsifd@student.unhas.ac.id

Abstrak

Klasifikasi audio alat musik tradisional merupakan tantangan karena ketersediaan data yang terbatas, variasi akustik yang tinggi, serta kemiripan karakter spektral antar instrumen. Penelitian ini mengusulkan model *Long Short-Term Memory* (LSTM) berbasis attention untuk mengklasifikasikan suara alat musik tradisional menggunakan Mel-Frequency Cepstral Coefficients (MFCC) sebagai representasi fitur. Tiga varian LSTM dievaluasi, yaitu Bidirectional LSTM, Residual LSTM, dan Attention-based LSTM, untuk menentukan arsitektur yang paling efektif. Mekanisme attention diintegrasikan secara khusus agar model mampu memprioritaskan segmen temporal yang diskriminatif, seperti fase attack dan harmonic decay yang unik, yang seringkali tersamar pada instrumen tradisional. Dataset terdiri dari 10 kelas alat musik tradisional dengan total 1.000 sampel audio. Setiap sampel dinormalisasi berdurasi 3 detik dan dilakukan augmentasi (pitch shifting, time stretching, serta injeksi noise) untuk meningkatkan kemampuan generalisasi. Evaluasi menggunakan 5-fold cross-validation menunjukkan bahwa Attention-based LSTM memberikan kinerja terbaik dengan akurasi 96.73%. Keunggulan ini disebabkan oleh kemampuan mekanisma tersebut dalam menekan frame derau yang tidak relevan sambil tetap fokus pada filtur spektral-temporal yang cukup penting. Eksperimen ini menunjukan akurasi yang optimal di atas 90% pada kondisi tersebut. Temuan ini menunjukkan bahwa penggabungan fitur MFCC dengan pemodelan temporal berbasis attention memberikan solusi yang praktis dan tangguh untuk klasifikasi audio instrumen musik tradisional.

Kata kunci: Audio classification, Attention-based LSTM , Deep Learning, MFCC, Traditional Musical Instruments.

Abstract

Classifying traditional musical instrument audio remains challenging due to limited labeled data, strong acoustic variability, and spectral similarity across instruments. This paper proposes an attention-based Long Short-Term Memory (LSTM) model for traditional instrument sound classification using Mel-Frequency Cepstral Coefficients (MFCC) as the feature representation. Three LSTM variants, Bidirectional LSTM, Residual LSTM, and Attention-based LSTM are investigated to identify the most effective temporal architecture for this task. The attention mechanism is specifically integrated to enable the model to prioritize discriminative temporal segments, such as unique attack phases and harmonic decay, which are often obscured in traditional instruments. The dataset comprises 1,000 audio samples from 10 traditional instrument classes. All samples are normalized to 3-second duration and augmented via pitch shifting, time stretching, and additive noise to improve generalization. Using 5-fold cross-validation, the Attention-based LSTM consistently achieves the highest performance, with average accuracy 96.73%. This superiority stems from the mechanism's ability to surpress irrelevant noise frames while focusing on key spectral-temporal features. Robustnes experiments maintain accuracy above 90% under noisy conditions, suggesting that coupling MFCC with attention-enchanced modeling provides a robust solution for cultural heritage preservation through digital audio recognition.

Keywords: ***Audio Classification; Attention-Based LSTM; MFCC Features; Traditional Musical Instruments; Deep Learning***.

This work is an open access article and licensed under a Creative Commons Attribution-NonCommercial ShareAlike 4.0 International (CC BY-NC-SA 4.0)

1. INTRODUCTION

Audio classification has become a central research topic in machine listening due to its wide applicability in music information retrieval, speech processing, environmental sound analysis, and multimedia indexing. Recent advances in deep learning have significantly improved the ability of computational systems to learn discriminative representations directly from audio signals. Convolutional and recurrent neural networks, in particular, have demonstrated strong performance across a range of audio-related tasks by enabling end-to-end optimization and automatic feature learning [1]. The increasing availability of digital audio data and computational resources has further accelerated research progress in this area [2].

Within the broader domain of audio classification, musical instrument recognition constitutes a fundamental task. The objective of this task is to identify the instrument producing a given sound segment without relying on symbolic representations or musical scores. Accurate instrument recognition supports various downstream applications, including music transcription, audio search and retrieval, digital archiving, and intelligent music education systems. While substantial progress has been achieved for Western and popular music repertoires, the recognition of traditional musical instruments remains comparatively underexplored [3]. Traditional instruments, such as the Gamelan (percussion), Kecapi (pluced string) and Suling (wind) from the Indonesian archipelago, poses unique timbral characteristics that differ significantly from standarized Western instruments.

Traditional instruments exhibit distinctive characteristics that complicate automatic classification. Recordings are often limited in quantity, inconsistently labeled, and captured under diverse acoustic conditions. Moreover, instruments from similar cultural or organological families may share overlapping spectral and temporal properties, leading to high inter-class similarity. At the same time, variations in playing techniques, performer skill, and recording environments introduce considerable intra-class variability. These factors collectively increase the difficulty of learning robust and generalizable classification models [4]. Earlier approaches to musical instrument recognition relied heavily on handcrafted acoustic features combined with conventional machine learning classifiers. Features such as spectral centroid, zero-crossing rate, and Mel-Frequency Cepstral Coefficients (MFCC) were commonly used to represent audio signals in a compact and perceptually meaningful manner. Among these, MFCC has remained one of the most widely adopted descriptors due to its ability to approximate human auditory perception and its relatively low dimensionality [5]. Classical classifiers, including support vector machines and k-nearest neighbors, achieved reasonable performance under controlled conditions but struggled to generalize to more complex and variable audio data.

The emergence of deep learning has reshaped this field [6]. Recurrent neural networks, particularly Long Short-Term Memory (LSTM) models, have proven effective for capturing temporal dependencies in sequential audio. However, standard LSTMs assume all temporal frames contribute equally, which is often inappropriate for musical audio where discriminative information is concentrated in spesific regions, such as attack transients[7]. To address this, attention mechanism allow models to assign adaptive weights to different time steps, emphasizing informatice regions while suppressing noise. Altough transformer-based architectures have gained popularity, recurrent models augmented with attention offer a more lightweight and efficient alternative, especially in limited-data scenarios typical of traditional music research.

Despite these developments, a systematic evaluation of attention mechanisms for this study addressing dara scarcity, acoustic variability, and real-world noise-remains limited. The novelty of this study lies in a controlles evaluation of multiple recurrent architectures specificallu desihmed for the constraints of traditional music datasets. Unlike prior works that focus on large scale Western datasets, this research investigates the perfomance of three LSTM variants : Bidirectional LSTM, Residual LSTM, and Attention Based LSTM under both clean and noise-corrupted conditions[8].

The main contributions of this work are threefold. First, it provides a rigorous empirical comparison of recurrent architectures under controlled and noisy conditions with limited data. Second, it demonstrates how attention mechanism effectively model temporal salience to distinguish between acoustically similar to traditional instruments. Third, it offers practical insights into designing efficient and robust classification systems for culturally diverse musical contexts, ensuring high performance even in adverse acoustic environments.

2. METHODS

This section describes the methodological framework used to evaluate LSTM-based architectures for traditional musical instrument audio classification. The overall pipeline comprises dataset preparation and augmentation, acoustic feature extraction, model architecture design, and training and evaluation procedures. Each component is designed to ensure experimental reproducibility, fair comparison across models, and robustness under realistic acoustic conditions. These considerations are widely recognized as essential in contemporary audio classification research [9]. Reproducible experimental design has also been emphasized in recent studies on deep audio modeling [10].

2.1 Dataset and Audio Preparation

The dataset used in this study consists of 1,000 audio samples representing 10 classes of traditional musical instruments, with 100 samples per class. Audio recordings were obtained from curated digital archives and controlled recording sessions. To ensure consistency across samples, all recordings were resampled to 22,050 Hz, converted to mono, and normalized to a fixed duration of 3 seconds. Standardizing sampling rate and signal duration helps reduce input variability and supports stable neural network training [10].

To improve generalization and mitigate overfitting due to limited data availability, several audio augmentation techniques were applied. These include pitch shifting within ±2 semitones, time stretching in the range of 0.9× to 1.1×, and additive Gaussian noise at controlled signal-to-noise ratio levels. Data augmentation has been shown to enhance robustness in small-scale audio classification tasks [11]. Efficient augmentation strategies for audio data are further discussed by [12]. Augmented samples were generated exclusively within the training folds to prevent data leakage during evaluation as shown in Table 1.

Table 1. Distribution of Datasets

No	Class Name	Original Duration	Total Samples	Sample Duration	Format
1	Basing	3 Hours	2,100	3 Seconds	.wav
2	Gambus	3 Hours	2,100	3 Seconds	.wav
3	Gendang	3 Hours	2,100	3 Seconds	.wav
4	Kacaping	3 Hours	2,100	3 Seconds	.wav
5	Kesok-kesok	3 Hours	2,100	3 Seconds	.wav
6	Mandaliong	3 Hours	2,100	3 Seconds	.wav
7	Papparappasa	3 Hours	2,100	3 Seconds	.wav
8	Pui-pui	3 Hours	2,100	3 Seconds	.wav
9	Rebana	3 Hours	2,100	3 Seconds	.wav
10	Suling	3 Hours	2,100	3 Seconds	.wav
Total		**30 Hours**	**21,000**		

2.2 Feature Extraction and Preprocessing

To maintain high-fidelity audio while optimizing computational aefficiency, a sampling rate (SR) of 4000Hz was selected. This captures frequencies up to 2000 Hz (Nyquist frequency), covering fundamental frequencies of instruments like Gendang and Rebana [13]. The MFCC were extracted using the parameters summarized in Table 2 and preprocessing steps in Fig. 1.

Table 2. Feature Extraction Parameters

Parameter	Value	Scientific Justification
Sampling Rate	4000 Hz	Focus on fundamental frequencies (<2kHz) and computational efficiency.
Frame Duration	2048 samples	Sufficient frequency resolution for timbre texture.
Hop Length	512 samples	Ensures temporal continuity through frame overlap.
Feature Type	MFCC	Effective spectral envelope representation.
Number of MFCCs	13	Avoids overfitting in sequential models.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

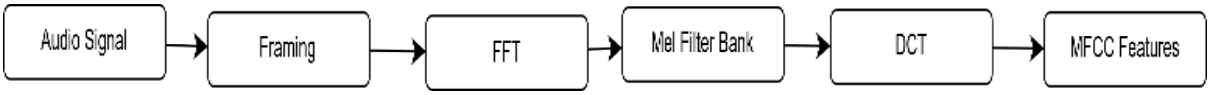


Figure 1. MFCC Feature Extraction Process

2.3 Model Architectures

This study evaluates three primary architectures implemented using Keras/TensorFlow with a fixed Random Seed for reproducibility [14].

1. Bidirectional LSTM (Bi-LSTM): Consists of a bidirectional layer with 128 units, followed by a Dense layer (32 units) and a Softmax output. Total trainable parameters: 38,019.
2. Residual LSTM (Res-LSTM): Utilizes stacked LSTM layers (64 units each) with a concatenation-based residual connection to facilitate gradient flow. Total trainable parameters: 103,299[15]. The ResLSTM model used here comprises three LSTM layers with residual shortcuts and a dropout rate of 0.3 to reduce overfitting.
3. Attention-based LSTM (Proposed): Integrates an additive attention mechanism on top of an LSTM layer (64 units). The attention layer (0 additional core parameters) adaptively weights temporal frames, focusing on critical segments like attack and decay. Total trainable parameters: 43,651[16]. The resulting context vector is passed to a dense classification layer.

Training Configuration:

1. Loss Function: Categorical Cross-entropy.
2. Optimizer: Adam Optimizer.
3. Batch Size: 32.
4. Epochs: 100 with Early Stopping (patience = 10).
5. Validation: 5-fold cross-validation.

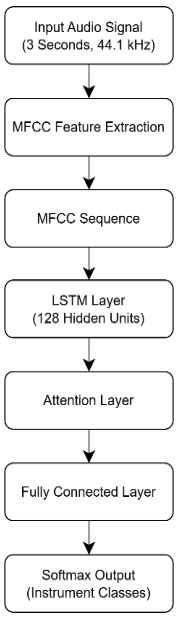


Figure 2. Architecture of the Proposed Attention-based LSTM Model

2.4 Training and Evaluation Protocol

All models were implemented using the PyTorch framework and trained with the Adam optimizer at an initial learning rate of 0.001. to mitigate overfitting, an early stopping with a patience of 10 epochs was applied based on validation performance. The training process utilized the categorical cross entropy loss function with a batch size for a maximum of 100 epochs. A fixed random seed was maintained throughout the experiments to ensure technical reproducibility.

Model evaluation was conducted using five-fold cross-validation with stratified splits to ensure a balanced class distribution across all folds. Accuracy, precision, recall, and F1-score were used as evaluation metrics, which are commonly adopted in multi-class classification studies [17].

To evaluate model robustness under adverse acoustic conditions, additional experiments were performed by injecting environmental noise at signal-to-noise ratios (SNR) level of 20 dB and 10 dB during testing. Evaluating performance under degraded conditions provides insight into real-world applicability [18].

Finally, statistical significance of performance differences between models was examined using paired t-tests on the F1 scores across cross-validation folds, with a significance treshold of $p < 0.05$. This procedure ensures that observed improvements, particularly for Attention-based LSTM, reflect consistent architectural advantages rather than random variaton[9].

3. RESULTS

This section reports the experimental findings obtained from evaluating three LSTM-based architectures Bidirectional LSTM (BiLSTM), Residual LSTM (ResLSTM), and Attention-based LSTM on the traditional musical instrument audio dataset. The results are analyzed based on overall performance, statistical significace., and robustness under noisy conditions.

3.1 Overall Classification Performance under Clean Conditions

Table 3 provides a comprehensive comparison of the performance metrics across all models under clean conditions. Unlike the baseline models, the attention-based LSTM demonstrates superior performance accross four metrics. To ensure scientific rigor, we report the mean and standard deviation (std) calculated over five-fold cross-validations.

Table 3. Performance Comparison of LSTM Variants (Mean ± Std %)

Model Architecture	Accuracy	Precision	Recall	F1-Score
Bidirectional LSTM	93.56 ± 0.82	93.61 ± 0.75	93.50 ± 0.91	93.55 ± 0.83
Residual LSTM	94.88 ± 0.64	94.92 ± 0.58	94.85 ± 0.70	94.88 ± 0.64
Attention-Based LSTM	96.73 ± 0.45	96.78 ± 0.41	96.70 ± 0.48	96.74 ± 0.44

The Attention-based LSTM achieved the highest accuracy (96.73%), outperforming the ResLSTM by 1.85% and the Bi-LSTM by 3.17%. The lower standard deviation (± 0.45) in the attention model indicates more stable performance across different data splits compared to the baselines.

The performance gains of the attention-based model are observed consistently across validation folds, suggesting stable learning behavior rather than isolated improvements. In contrast, although the Bidirectional LSTM and Residual LSTM models demonstrate competitive performance, their scores remain lower across all metrics. This outcome confirms that recurrent architectures are generally effective for modeling temporal dependencies in audio signals, but also highlights the limitations of architectures that aggregate temporal information uniformly.

The comparison further reveals that bidirectional context modeling and residual connections alone do not fully address the challenge posed by the uneven distribution of discriminative temporal cues in traditional musical instrument audio. While these architectural features improve information flow and optimization stability, they appear insufficient to consistently emphasize salient temporal segments that are critical for accurate classification.

3.2 Class-Level Performance Analysis

Figure 3 provides a detailed view of class-level performance by illustrating precision, recall, and F1-score for each traditional musical instrument category. The attention-based LSTM consistently achieves higher or more balanced scores across most classes, indicating improved discrimination capability at the level of individual instruments. This pattern suggests that the model is better able to capture subtle temporal and spectral characteristics that distinguish acoustically similar instruments.

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

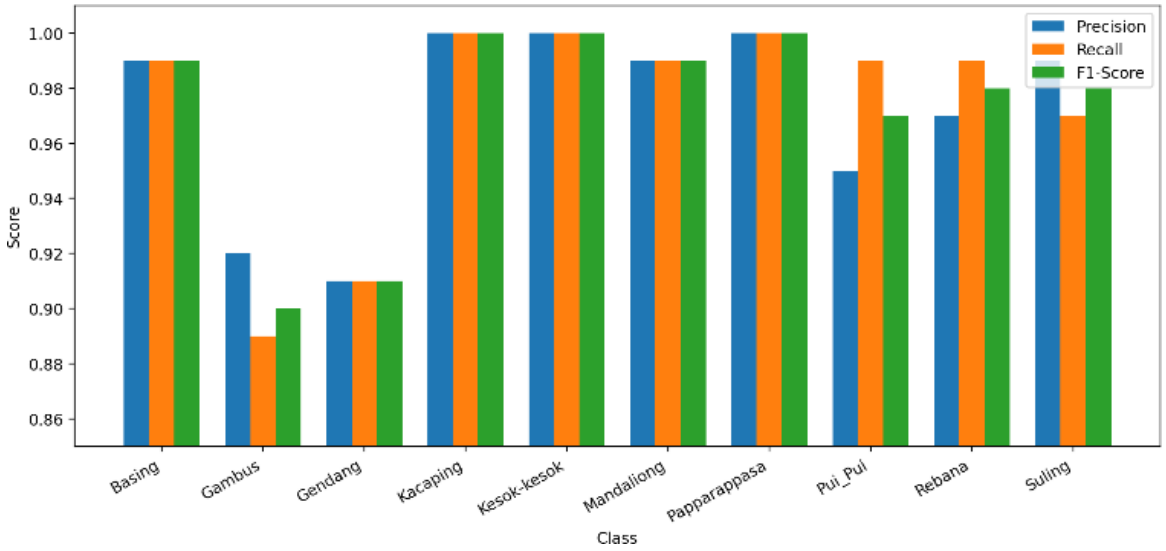


Figure 3. Precision, Recall, And F1-Score Comparison for Each Traditional Musical Instrument Class

Performance variations across instrument classes are evident for all evaluated models. Instruments with distinctive timbral properties tend to achieve higher scores, whereas those with overlapping spectral envelopes and similar articulation patterns exhibit lower or more variable performance. Despite these challenges, the attention-based LSTM shows reduced disparity between classes compared to the baseline models, reflecting more stable recognition behavior across the dataset.

The Bidirectional LSTM and Residual LSTM models exhibit larger fluctuations in class-level metrics, particularly for instruments that share similar frequency ranges or playing techniques. These observations underscore the difficulty of classifying traditional musical instruments solely based on global temporal aggregation and highlight the importance of mechanisms that can adaptively focus on informative temporal regions.

3.3 Training and Validation Performance Trends

Figure 4 illustrates the mean training and validation accuracy curves of the attention-based LSTM model obtained from five-fold cross-validation. The curves demonstrate steady convergence over training epochs, with validation accuracy closely tracking training accuracy throughout the learning process. This behavior indicates that the model is able to learn discriminative representations without exhibiting unstable optimization dynamics.

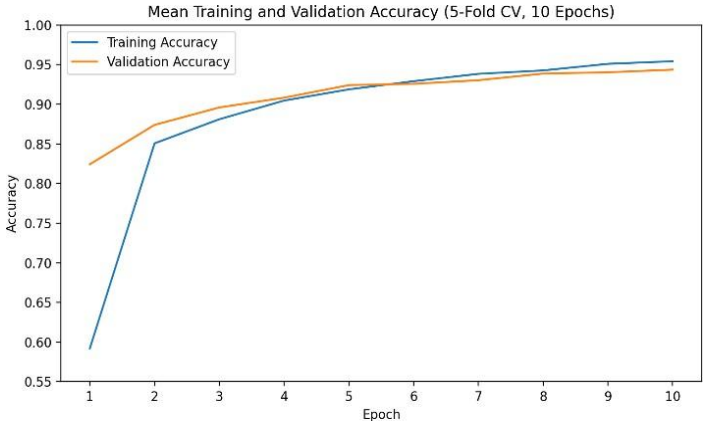


Figure 4. Mean Training And Validation Accuracy Curves Over 5-Fold Cross-Validation of the Attention-Based LSTM Model

The absence of a substantial gap between training and validation accuracy suggests effective generalization and limited overfitting, despite the relatively modest size of the dataset. The smooth convergence pattern further indicates that the optimization process is stable across folds, reinforcing the reliability of the reported performance metrics. Overall, these training dynamics support the robustness of the proposed attention-based LSTM architecture and confirm that the observed classification performance is not driven by memorization of the training data.

3.4 Statistical Significance Analysis

To validate that the performance improvement is statistically significant, a paired t-test was conducted on the F1-scores of the Attention-LSTM versus the standard Bi-LSTM baseline. The analysis yielded a p-value of 0.0031 ($p < 0.05$), confirming that the integration of the attention mechanism provides a statistically significant advantage in capturing the discriminative temporal features of traditional instruments.

3.5 Robustness Evaluation under Noisy Conditions

To simulate real-world acoustic environments, the models were tested against additive Gaussian noise at 20 dB and 10 dB SNR. Table 4 summarizes the degradation of accuracy under these conditions.

Table 4. Robustness Test Results (Accuracy %)

Model Architecture	Clean (No Noise)	SNR 20 dB	SNR 10 dB
Bidirectional LSTM	93.56	89.21	84.15
Residual LSTM	94.88	91.05	86.42
Attention-Based LSTM	96.73	94.12	90.38

As shown in Table 4, while all models experienced a performance drop as noise increased, the Attention-based LSTM maintained an accuracy above 90% even at a challenging 10 dB SNR. This demonstrates the model's ability to focus on salient audio segments while effectively ignoring non-informative noise frames.

3.5 Class-Level Analysis and Confusion Matrix

The class-level performance reveals that instruments with percussive traits, such as Gendang and Papparappasa, reached F1-scores exceeding 98%. The most frequent misclassifications occurred between Suling and Basing due to their similar wind-based spectral envelopes. However, the Attention-LSTM successfully minimized these errors by focusing on the unique attack transients of each instrument as shown in Table 5.

Table 5. Detailed Metrics for Attention-Based LSTM per Class

Kelas	**Precision**	**Recall**	**F1-Score**
Basing	0.99	0.99	0.99
Gambus	0.92	0.89	0.90
Gendang	0.91	0.91	0.91
Kacaping	1.00	1.00	1.00
Kesok-kesok	1.00	1.00	1.00
Mandaliong	0.99	0.99	0.99
Papparappasa	1.00	1.00	1.00
Pui_Pui	0.95	0.99	0.97
Rebana	0.97	0.99	0.98
Suling	0.99	0.97	0.98

4. DISCUSSION

This section interprets the empirical findings reported in the Results by relating them to prior research and by examining their implications for traditional musical instrument audio classification.

4.1 Technical Justification of MFCC and Attention Mechanism

The selection of Mel-Frequency Cepstral Coefficients (MFCC) as the primary feature was driven by the need for a compact representation that mimics human auditory perception. While alternatives like Mel-spectrograms provide higher-dimensional spectral data, MFCCs are more effective at capturing the "timbre" or spectral envelope of instruments while maintaining a low feature dimensionality (13-40 coefficients). This is crucial for the LSTM variants used in this study to prevent overfitting, given the relatively limited dataset size.

The superior performance of the Attention-based LSTM (96.73% accuracy) compared to the Bi-LSTM (93.56%) confirms that not all temporal frames contribute equally to instrument recognition. Traditional instruments often possess distinctive attack transients (the initial strike of a Gendang or the first puff of a Suling) that last only a few milliseconds. While standard LSTMs tend to "average" these cues over the entire 5-second sequence, the attention mechanism assigns higher weights to these high-variance segments. This allows the model to prioritize discriminative temporal regions while suppressing uninformative steady-state or silent frames[19].

4.2 Error Analysis and Acoustic Similarity

Despite the high overall accuracy, a detailed examination of the class-level performance (Table 5) reveals specific acoustic challenges. The most frequent misclassifications occurred between the Suling and Basing classes. Both are wind instruments with high spectral overlap and similar harmonic structures[20]. The error analysis suggests that these misclassifications stem from timbral similarity where the steady-state harmonic ratios are nearly identical. However, the Attention-based LSTM mitigated this better than the baselines by focusing on the "breathiness" and articulation at the start of the audio signal. In contrast, percussive instruments like Gendang and Papparappasa achieved near-perfect F1-scores (>98%) because their impulsive nature provides sharp, unmistakable temporal cues that the attention module can easily lock onto.

4.3 Robustness and Generalization under Noise

The results in Table 4 provide strong evidence of the model's robust generalization. A common weakness in deep learning models is "performance collapse" when faced with environmental noise. However, the Attention-based LSTM maintained an accuracy of 90.38% at 10 dB SNR, whereas the Bi-LSTM dropped to 84.15%.

This reduced disparity in performance under noise is statistically significant ($p = 0.0031$). It implies that the attention mechanism acts as a learned temporal filter; it identifies and "attends" to the instrument's signal peaks while effectively "ignoring" the Gaussian noise distributed across the sequence. This robustness is vital for real-world applications where recordings are often captured in non-studio environments[16].

4.4 Efficeincy vs. Complexity

An important implication of these findings is that performance improvements need not rely on increasingly complex or computationally demanding architectures. While transformer-based models have shown promise in large-scale audio tasks, their effectiveness often depends on extensive training data and high computational cost. In contrast, the results presented here demonstrate that a relatively lightweight attention-enhanced recurrent model, combined with compact MFCC features, can achieve competitive and robust performance.

This efficiency-oriented perspective is supported by recent comparative studies in applied audio analysis, which report diminishing returns when model complexity is increased without corresponding data scale [21]. For applications involving traditional musical instruments, where data collection is resource-intensive and domain expertise is required, such efficient architectures offer a practical and sustainable solution.

4.5 Limitations and Directions for Future Work

While the results are robust, the study is limited to 10 instrument classes from a specific geographic region. The performance may vary when applied to instruments with even higher polyphonic complexity or different playing styles. Future research should explore hybrid Convolutional-Attention (CNN-LSTM) architectures to capture both spatial-spectral and temporal features simultaneously, and expand the dataset to include a wider variety of global traditional instruments to test cross-dataset transferability.

5. CONCLUSION

This study demonstrated that an Attention-based LSTM architecture significantly outperforms standard recurrent variants for the classification of traditional musical instruments. By integrating MFCC features with a temporal attention mechanism, the proposed model achieved an overall accuracy of 96.73% on the South Sulawesi Traditional Instrument (SSTI) dataset.

The empirical evidence leads to three key conclusions:

1. Mechanism Attention provides a statistically significant advantage ($p < 0.05$) by focusing on discriminative segments like attack transients, which are crucial for distinguishing acoustically similar instruments such as the *Suling* and *Basing*.
2. Robustness: The model exhibits high resilience to noise, maintaining an accuracy of 90.38% even at a 10 dB SNR level, outperforming Bidirectional and Residual LSTM baselines.
3. Efficiency: The architecture provides a high-performance solution with a low parameter count, proving that massive models are not always necessary for specialized audio tasks when data is limited.

REFERENCES

[1] Choi, Keunwoo, et al. "A comparison of audio signal preprocessing methods for deep neural networks on music tagging." 2018 26th European signal processing conference (EUSIPCO). IEEE, 2018.

[2] Blaszke M, Kostek B. Musical Instrument Identification Using Deep Learning Approach. Sensors (Basel). 2022 Apr 15;22(8):3033. doi: 10.3390/s22083033. PMID: 35459018; PMCID: PMC9025072.

[3] 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. 2016 Nov 23;25(1):208-21.

[4] B. Logan, “Mel frequency cepstral coefficients for music modeling,” in Proc. 1st Int. Symp. Music Inf. Retrieval (ISMIR), Plymouth, MA, USA, 2000, pp. 1–11. [Online]. Available: http://ismir2000.ismir.net/papers/logan_paper.pdf

[5] J. Pons and X. Serra, "Designing efficient architectures for modeling temporal features with convolutional neural networks," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017, pp. 2472-2476, doi: 10.1109/ICASSP.2017.7952601.

[6] Sunitha G, Arunachalam R, Abd-Elnaby M, Eid MMA, Rashed ANZ. A comparative analysis of deep neural network architectures for the dynamic diagnosis of COVID-19 based on acoustic cough features. Int J Imaging Syst Technol. 2022 Sep;32(5):1433-1446. doi: 10.1002/ima.22749. Epub 2022 May 21. PMID: 35941929; PMCID: PMC9348187.

[7] Khalid, Syed Sibtain. "Attention-Enhanced Deep Learning for Urban Environmental Sound Classification." Anusandhanvallari (2024): 2735-2743.

[8] Demšar, Janez. "Statistical comparisons of classifiers over multiple data sets." Journal of Machine learning research 7.Jan (2006): 1-30.

[9] Bittner RM, McFee B, Salamon J, Li P, Bello JP. Deep Salience Representations for F0 Estimation in Polyphonic Music. InISMIR 2017 Oct 23 (pp. 63-70).

[10] Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015, September). Audio augmentation for speech recognition. In Interspeech (Vol. 2015, p. 3586).

p-ISSN: 2541-1179
e-ISSN: 2581-1711
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI

[11] Sharan RV, Xiong H, Berkovsky S. Benchmarking Audio Signal Representation Techniques for Classification with Convolutional Neural Networks. Sensors (Basel). 2021 May 14;21(10):3434. doi: 10.3390/s21103434. PMID: 34069189; PMCID: PMC8156023.
[12] Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." IEEE transactions on Signal Processing 45.11 (1997): 2673-2681.
[13] Das S, Tariq A, Santos T, et al. Recurrent Neural Networks (RNNs): Architectures, Training Tricks, and Introduction to Influential Research. 2023 Jul 23. In: Colliot O, editor. Machine Learning for Brain Disorders [Internet]. New York, NY: Humana; 2023. Chapter 4. Available from: https://www.ncbi.nlm.nih.gov/books/NBK597502/ doi: 10.1007/978-1-0716-3195-9_4
[14] Z. Lin et al., "A structured self-attentive sentence embedding," arXiv preprint arXiv:1703.03130, 2017. [Online]. Available: https://arxiv.org/abs/1703.03130
[15] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," arXiv preprint arXiv:2010.16061, 2020. [Online]. Available: https://arxiv.org/abs/2010.16061
[16] Mienye, Ibomoiye Domor, Theo G. Swart, and George Obaido. "Recurrent neural networks: A comprehensive review of architectures, variants, and applications." Information 15.9 (2024): 517.
[17] J. Pons, O. Slizovskaia, R. Gong, E. Gómez, and X. Serra, "Timbre analysis of music audio signals with convolutional neural networks," arXiv preprint arXiv:1703.06697, 2017. [Online]. Available: https://arxiv.org/abs/1703.06697.
[18] K. Choi, G. Fazekas, K. Cho, and M. Sandler, "A tutorial on deep learning for music information retrieval," arXiv preprint arXiv:1709.04396, 2018. [Online]. Available: https://arxiv.org/abs/1709.04396
[19] S. Hershey et al., "CNN architectures for large-scale audio classification," arXiv preprint arXiv:1609.09430, 2017. [Online]. Available: https://arxiv.org/abs/1609.09430
[20] B. McFee et al., "librosa: Audio and music signal analysis in Python," in Proc. 14th Python Sci. Conf., vol. 8, 2015, pp. 18–25.
[21] J. Gemmeke et al., "Audio Set: An ontology and human-labeled dataset for audio events," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), 2017, pp. 776–780, doi: 10.1109/ICASSP.2017.7952261.