SCIENCE AND INFORMATION TECHNOLOGY
SINTECH
JOURNAL

SINTECH JOURNAL | ISSN 2598-7305 | E-ISSN 2598-9642
Vol. 8 No 2 – Agustus 2025 | https://s.id/sintechjournal
DOI: https://doi.org/10.31598
Publishing: Prahasta Publisher

Analysis Of Neural Network Architectures For Syllable-Based Voice Recognition In Indonesian

Deni Sutendi Kartawijaya [1], Tjong Wan Sen [2]

[1,2]Magister of Informatic Study Program, Faculty of Computing, President University
Jababeka Education Park, Jl. Ki Hajar Dewantara, Bekasi, Indonesia

e-mail: deni.kartawijaya@student.president.ac.id [1], wansen@president.ac.id [2]

Received : December, 2024	Accepted : August, 2025	Published : August, 2025

Abstract

Nowadays, speech recognition technology is widely used in various technology platforms. But there are still only a few numbers of researchs on speech recognition in Indonesian syllable recognition. The main goal of the research is to implement the combination of several deep learning techniques to get the best Model-Based Recognition Systems for Indonesian syllable recognition. Due to the limited of time, current research was conduted to get the best knowledge on how to process syllable voice recognition in Indonesian using 1-D array data using 3 deep learning techniques such as Artificial Neural Networks (ANN), Long Short-Term Memory Networks (LSTM), and Convolutional Neural Networks (CNN). Based on those situations, this study focuses on syllable-based voice recognition in Indonesian using 1D array data that evaluates and compares the performance ANN, LSTM, and CNN, to determine their effectiveness in recognizing syllables within voice data. The dataset of voice recordings was conducted manually. The labeling process was conducted by manually segmenting the 1D array form of the voice data to get the most accurate label. Each syllable was divided into 3 parts with the same size (1024 time-based array data). At the beginning, there were 400 voice recordings collected, but due to the limited of time for the task submission, 10 voice recordings were processed resulting in 309 unique syllable parts across 60 classes. Each architecture was evaluated for their accuracy. The results indicate significant differences in model performance, with CNN demonstrating superior capabilities in capturing sequential dependencies inherent in syllabic speech data. Based on the experiments, the CNN model is the best model to process the Indonesian syllable classification with 99.86% accuracy, followed by LSTM and ANN with 99.03% and 91.91% accuracy respectively. This study may contribute to the next process for Indonesian voice recognition as a basis to conduct another research by combining these models to get the best result.

Keywords: *voice recognition, syllable recognition, ANN, LSTM, CNN, deep learning, Indonesian*

Abstrak

Saat ini, teknologi pengenalan suara banyak digunakan di berbagai platform teknologi. Namun, masih sedikit penelitian tentang pengenalan suara dalam pengenalan suku kata bahasa Indonesia. Tujuan utama dari penelitian ini adalah untuk mengimplementasikan kombinasi beberapa teknik deep learning untuk mendapatkan Sistem Pengenalan Berbasis Model terbaik untuk pengenalan suku kata bahasa Indonesia. Karena keterbatasan waktu, penelitian saat ini dilakukan untuk mendapatkan pengetahuan terbaik tentang cara memproses pengenalan suara suku kata dalam bahasa Indonesia menggunakan data array 1-D menggunakan 3 teknik deep learning seperti Artificial Neural Networks (ANN), Long Short-Term Memory Networks (LSTM), dan Convolutional Neural Networks (CNN). Berdasarkan situasi tersebut, penelitian ini berfokus pada pengenalan suara berbasis suku kata dalam bahasa Indonesia menggunakan

data array 1D yang mengevaluasi dan membandingkan kinerja ANN, LSTM, dan CNN, untuk menentukan efektivitasnya dalam mengenali suku kata dalam data suara. Kumpulan data rekaman suara dilakukan secara manual. Proses pelabelan dilakukan dengan melakukan segmentasi manual bentuk array 1D dari data suara untuk mendapatkan label yang paling akurat. Setiap suku kata dibagi menjadi 3 bagian dengan ukuran yang sama (1024 data array berbasis waktu). Pada awalnya, ada 400 rekaman suara yang dikumpulkan, tetapi karena keterbatasan waktu untuk penyerahan tugas, 10 rekaman suara diproses sehingga menghasilkan 309 bagian suku kata unik di 60 kelas. Setiap arsitektur dievaluasi untuk akurasinya. Hasilnya menunjukkan perbedaan yang signifikan dalam kinerja model, dengan CNN menunjukkan kemampuan yang unggul dalam menangkap dependensi sekuensial yang melekat pada data ucapan suku kata. Berdasarkan percobaan, model CNN adalah model terbaik untuk memproses klasifikasi suku kata bahasa Indonesia dengan akurasi 99,86%, diikuti oleh LSTM dan ANN dengan akurasi masing-masing 99,03% dan 91,91%. Penelitian ini dapat berkontribusi pada proses selanjutnya untuk pengenalan suara bahasa Indonesia sebagai dasar untuk melakukan penelitian lain dengan menggabungkan model-model ini untuk mendapatkan hasil terbaik.

Kata Kunci: *pengenalan suara, pengenalan suku kata, ANN, LSTM, CNN, deep learning, bahasa Indonesia*

1. INTRODUCTION

Speech recognition systems have become critical for modern applications, ranging from virtual assistants to automated transcription. Despite advancements, most research targets resource-rich languages like English. In contrast, Bahasa Indonesia, with its consistent syllable structures, remains underrepresented.

Despite significant advancements in speech recognition, most systems are optimized for resource-rich languages like English. Bahasa Indonesia lacks publicly available syllable-based datasets, making it challenging to develop accurate voice-enabled technologies. This study addresses this gap by creating a manually labeled dataset and evaluating deep learning models for Indonesian syllable recognition, contributing to the development of more inclusive speech recognition systems.

The unique phonetic and syllabic patterns of Bahasa Indonesia necessitate specialized models. Syllable-based recognition, as opposed to word- or phoneme-level approaches, offers advantages in low-resource linguistic contexts.

Despite significant progress in speech recognition, the performance of such systems varies greatly across languages. With its distinct linguistic features, Indonesian remains underrepresented in global research on voice recognition. Existing systems often fail to capture the nuances of syllables in Bahasa Indonesia, leading to suboptimal recognition performance. This problem necessitates a focused exploration of advanced neural network architectures—specifically, Artificial Neural Networks (ANN), Long Short-Term Memory Networks (LSTM), and Convolutional Neural Networks (CNN)—to evaluate their suitability for syllable recognition tasks. To get the correct form of datasets, the audio raw data could be formed in 1-dimensional array data, Fourier Transform Data, MFCC spectrum data, etc. The main goal of the research is to implement the combination of several deep learning techniques to get the best Model-Based Recognition Systems for Indonesian syllable recognition. Due to the limited of time, current research was conduted to get the best knowledge on how to process syllable voice recognition in Indonesian using 1-D array data using 3 deep learning techniques such as Artificial Neural Networks (ANN), Long Short-Term Memory Networks (LSTM), and Convolutional Neural Networks (CNN).

Because of thet, in this study, the research was conducted by manually segmented the raw audio data into 1D array data to be used as input for deep learning methods, including the ANN, LSTM, and CNN models. Despite advancements in voice recognition, most research focuses on resource-rich languages like English, leaving languages like Indonesian underexplored. Research often emphasizes phoneme- or word-level analysis, overlooking the potential of syllable-based approaches for languages with consistent syllable patterns like Indonesian. Few annotated syllable-level datasets exist for Bahasa Indonesia, requiring reliance on manual segmentation, which is labor-intensive but necessary for accuracy in low-resource contexts. The comparative effectiveness of ANN, LSTM, and CNN for syllable recognition in Indonesian remains unexamined, despite their prominence in speech recognition. Automated segmentation tools are often inadequate for capturing the nuances of Bahasa Indonesia syllables, necessitating manual segmentation. This study addresses these gaps by utilizing a manually

segmented Indonesian syllables dataset to evaluate ANN, LSTM, and CNN, contributing to the development of effective Indonesian syllable-based voice recognition.

To address the gaps, this study compares three neural network architectures—ANN, LSTM, and CNN—on a manually segmented dataset.

The objectives of this study are:

1) To manually preprocess and segment Indonesian speech into syllable-level datasets.
2) To evaluate ANN, LSTM, and CNN architectures for syllable-based classification.
3) To compare their performance using accuracy metrics.

2. RELATED WORK

Speech recognition research has evolved with deep learning models, including ANN, LSTM, and CNN:

1) ANN: Basic feedforward networks, while effective for simple tasks, struggle with sequential dependencies in speech signals.
2) LSTM: A variant of RNNs, LSTMs are designed to model temporal sequences effectively, making them ideal for speech data.
3) CNN: Originally applied to images, CNNs can extract features from audio waveforms, efficiently capturing local dependencies.

Several studies highlight the integration of deep learning with speech recognition systems. (Khdier et al., 2021) demonstrated CNN's robustness in noisy environments using raw waveforms. (Guan et al., 2024) combined deep learning with language models to improve speech accuracy, (Suyanto et al., 2021) focused on syllable-level analysis in Indonesian using BiLSTM-CNN models.

Recent advancements in speech recognition, such as Transformer-based models (e.g., Wav2Vec2), have demonstrated state-of-the-art performance in end-to-end speech tasks. However, these models require extensive training data, making them less feasible for syllable-based recognition in low-resource languages. Future research will explore fine-tuning pre-trained Transformer models for Indonesian syllable recognition to determine their effectiveness compared to CNN, LSTM, and ANN.

3. *METHODOLOGY*

This research proposes to find a way how to explore Artificial Neural Networks (ANN), Long Short-Term Memory Networks (LSTM), and Convolutional Neural Networks (CNN) on syllable-based voice recognition in Indonesian. The study involves manually segmenting audio recordings into syllables, dividing each syllable into three equal parts, and training the data using deep learning models to classify these segments into unique syllable-part classes. All processes are done in the Google Colab environment.

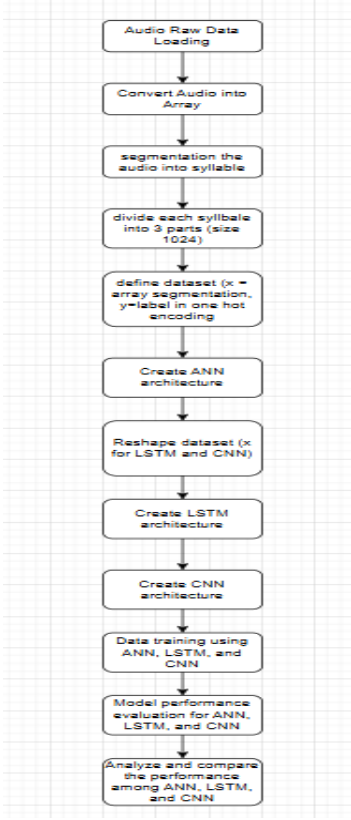

Figure 1. Flowchart of the business process

Based on Figure 1. The first step is audio raw data loading. After that, convert the audio data into array data. Then segmenting the audio into syllable. After that, divide each syllable into 3 parts equally (1024 array data). The next step is to define dataset (x = array syllable segmentation, y = label in one hot encoding). And then create 3 model architectures (ANN, LSTM, and CNN). After that, do data training using all the dataset as data training and validation for all 3 models. Finally calculate the model performance among those 3 models and analyse and comparing all the models accuracy.

3.1 Dataset preparation

At the beginning, there were 400 voice recordings collected. Due to a limited time for the task submission, there are only 10 voice recordings in Bahasa Indonesia were processed, featuring phrases like "Selamat pagi semuanya". Each recording was manually segmented into syllables using Audacity software. Each syllable was further divided into three equal parts of 1024 time-based

array samples, resulting in 309 unique syllable parts across 60 classes, as shown in Table 1.

Due to dataset limitations, there is a risk of overfitting, especially in deep learning models like CNN and LSTM. To address this, we applied hyperparameter tuning to prevent overfitting. Additionally, future work will explore data augmentation techniques such as time-stretching and pitch shifting to enhance generalizability. And also consider to use dropout regularization and early stopping.

Table 1: Example dataset

File name	syllables
Voice1.wav	se, la, mat, pa, gi, se, mu, a, nya
Voice2.wav	se, la, mat, si, ang, ba, pak, I, bu
Voice3.wav	se, la, mat, so, re, te, man, te, man
Voice4.wav	se, la, mat, dan, mim, pi, in. dah
Voice5.wav	a, pa, ka, bar, de, ngan, ha, ri, i, ni
Voice6.wav	se, mo, ga, ha, ri, i, ni, me, nye, nang, kan
Voice7.wav	sam, pai, jum, pa, di, la, in, wak, tu
Voice8.wav	Se, nang, ber, te, mu, de, ngan, an, da, la, gi
Voice9.wav	ha, ri, i, ni, cu, a, ca, nya, sa, ngat, ce, rah
Voice10.wav	sa, ya, i, ngin, mem, be, li, sa, tu, bu, ah, a, pel

Manual segmentation was chosen to ensure high labeling accuracy, which is critical for benchmarking syllable-based models. However, we acknowledge that manual segmentation is time-consuming and not scalable for large datasets. Future research will explore automated segmentation techniques, such as Forced Alignment or neural network-based segmentation, to improve scalability while maintaining labeling accuracy.

With Audacity software as shown in Figure 2, it can be defined the boundaries of each syllable by listening sharply part by part, so that the accurately is high to segment the signal into parts of each syllable. By using this method, further step is segmenting from the array sound.

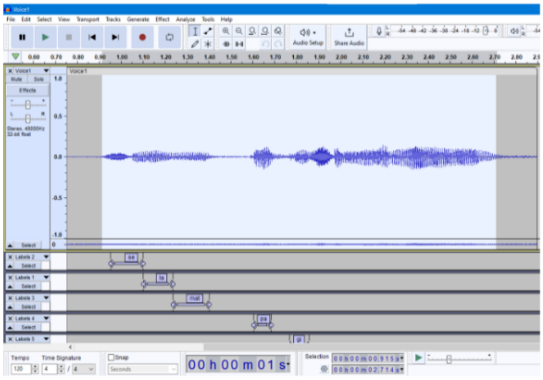

Figure 2. The signal of voice1.wav in audacity.

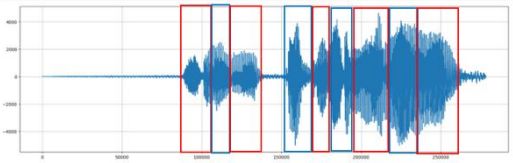

Figure 3. Mapping each syllable from the signal of voice1.wav form

As shown in Figure 3 as an example, the part of voice1 (ad1) was segmented from the array segment of 88000 – 108000 as syllable ‘se’ from the voice “selamat pagi semuanya”.

```
ad1[85000:110000]
```

Show hidden output

```
#se
plt.figure(figsize=[4,5])
plt.title('se')  # Adding title
plt.grid(True)    # Adding grid
plt.plot(ad1[88000:108000])
se=ad1[88000:108000]
```

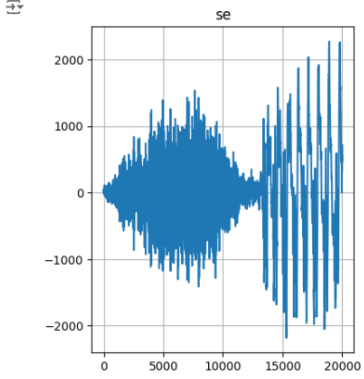

Figure 4. Segmenting syllable “se” from of voice1 (ad1) from 88000 – 108000

3.2 Neural Network Architectures

Three models were implemented using TensorFlow/Keras:

3.2.1 ANN

1D array (1024 features) was used for the input layer. For the Architecture using 1 input layer, 2 hidden layers (ReLU activation), and 1 output layer (Softmax/Sigmoid). And for the loss function using

Categorical Crossentropy (Softmax) and Binary Crossentropy (Sigmoid).

3.2.2 LSTM

1D array data (1024 feature) was reshaped to 3D tensor (samples, timesteps=1024, features=1) in the input layer. While the architecture consist of two LSTM layers followed by a dense output layer.

3.2.3 CNN

1D array data (1024 feature) was reshaped into 3D tensor (samples, timesteps=1024, features=1) in the input layer. While the architecture consist of Conv1D layers with kernel sizes 3 and 5, MaxPooling, and GlobalAveragePooling. And for the activation function used ReLU in hidden layers, and Softmax/Sigmoid in the output layer.

3.3 Training and Evaluation

The models were trained using Adam optimizer for up to 200 epochs with batch size 32.

Adam is widely used because it combines the benefits of AdaGrad (adaptive learning rate) and RMSprop (momentum-based updates).

The number of epochs depends on dataset size, complexity, and model convergence. Common references for training deep learning models suggest starting with 50–200 epochs for moderate-sized datasets.

Batch size 32 is a common choice as a trade-off between stability and computational efficiency. Evaluation was based on accuracy.

4. RESULT AND DISCUSSION

4.1 ANN Performance.

The ANN model achieved an accuracy of 90.61% with the Softmax activation and slightly improved with 91.91% using Sigmoid. Despite good performance, ANN struggled with capturing temporal dependencies, as shown in Figure 5 and Figure 6 below.

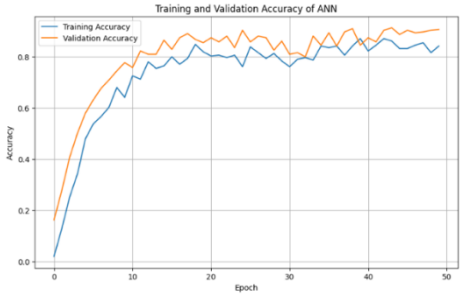

Figure 5. Training and Validation Accuracy Graph (ANN-Softmax)

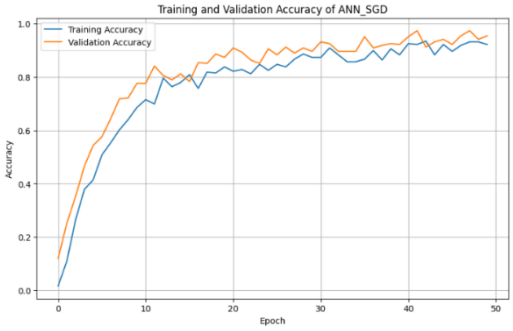

Figure 6. Training and Validation Accuracy Graph (ANN-Sigmoid)

4.2 LSTM Performance.

LSTM outperformed ANN, achieving **93.85% accuracy** using softmax activation function with 50 epochs, as shown in Figure 7. Its ability to model temporal dependencies contributed to better recognition of syllables.

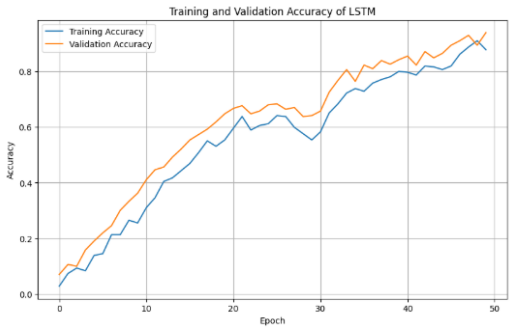

Figure 7. Training and Validation Accuracy Graph (LSTM-Softmax)

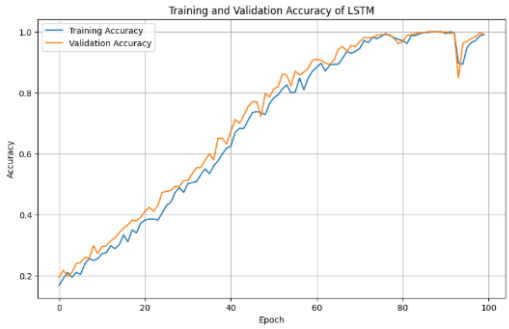

Figure 8. Training and Validation Accuracy Graph (LSTM-Sigmoid)

But, when using the Sigmoid activation function in LSTM, ss shown in Figure 8, with 50 epochs, the accuracy of predicting the syllable classification is around 22.65%, and needs more than 100 epochs to reach 99.05%.

4.3 CNN Performance.

As shown in Figure 9, we can observe that using the softmax activation function on the CNN model, with 200 epochs, the accuracy of predicting the syllable classification is 99.68%. while as shown in

Figure 10, we can observe that using the sigmoid activation function on the CNN model, with 200 epochs, the accuracy of predicting the syllable classification is 78.64%.

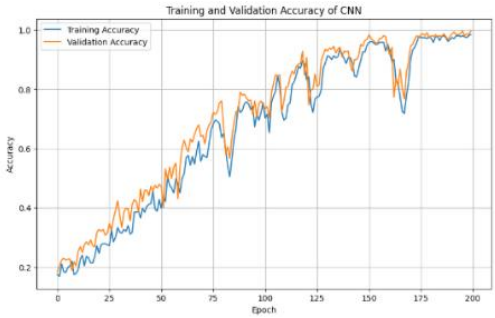

Figure 9. Training and Validation Accuracy Graph (CNN-softmax)

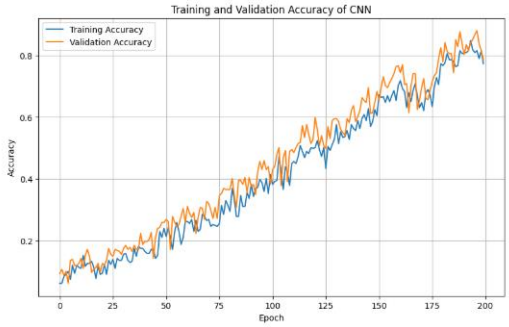

Figure 10. Training and Validation Accuracy Graph (CNN-sigmoid)

The CNN model demonstrated superior performance, achieving 99.86% accuracy. Its ability to extract hierarchical features from raw waveforms allowed it to outperform ANN and LSTM.

4.4 Summary Result

The accuracy of each models can be be seen on Table 2.

Table 2. Accuracy comparison

Model	Activation function	Number of Epoch	Accuracy
ANN	Softmax	50	90.61%
	Sigmoid	50	91.91%
LSTM	Softmax	50	**93.85%**
	Sigmoid	100	99.03%
CNN	Softmax	200	**99.68%**
	Sigmoid	200	78.64%

Reducing the complexity of the model (e.g., reducing the number of layers or units in a neural network) can also help prevent overfitting.

To prevent overfitting, actually there are several techniques can be employed, including early stopping, regularization, cross-validation, model complexity, data augmentation, and feature selection. While in this research, model complexity technique was conducted by doing several experiment during creating the mocel architecture and training process. By adding or subtracting layers or units in each model architecture to get the best accuracy result.

4.5 Discussion

Based on all the experiments as shown in Table 3, the iteration speed and accuracy across different models and training settings reveal distinct strengths and trade-offs. Among these, the LSTM model with a softmax activation function trained for 50 epochs achieved 93.85% accuracy, demonstrating its capability in sequential pattern recognition. This result was derived from the 1-dimensional array data used as the dataset, reinforcing LSTM's suitability for modeling temporal dependencies in syllable classification.

However, when considering the overall highest accuracy, the CNN model with a softmax activation function trained for 200 epochs reached 99.68% accuracy, surpassing both the ANN and LSTM models. This suggests that, given a longer training duration, CNNs are highly effective in feature extraction and classification. Nevertheless, this comes at the cost of increased computational demands and longer training times, as CNNs traditionally excel in 2D spatial data processing rather than 1D sequential data.

4.5.1 Comparison of Model Performance Based on Training and Validation Accuracy

As depicted in Figure 5 and Figure 6, the ANN model exhibited notable efficiency due to its simple architecture and faster training time. This efficiency makes it particularly useful for smaller or less complex datasets, where rapid convergence is prioritized over intricate feature extraction. The ANN model's simplicity also enhances its stability in generalizing moderately complex data, though it lacks the specialized feature extraction capabilities of CNNs or the sequential learning advantages of LSTMs.

On the other hand, the LSTM model excelled in processing sequential syllable parts due to its inherent ability to capture long-range dependencies in time-series data. This characteristic is crucial for syllable-based voice

recognition, where the temporal order of syllables significantly affects meaning and pronunciation. The LSTM model's 99.03% accuracy after just 50 epochs indicates its efficiency in leveraging sequential information to improve classification performance.

Meanwhile, the CNN model demonstrated robust feature extraction and strong resistance to noise, contributing to its superior accuracy of 99.68%. However, one of its primary limitations is its higher computational cost and longer convergence time. Unlike ANN and LSTM models, which required only 50 epochs to reach their peak accuracies (91.11% for ANN and 99.03% for LSTM), CNN required 200 epochs to achieve its highest accuracy. This extended training time is likely due to CNN's inherent structure, which processes data spatially, making it computationally intensive when applied to 1D array data instead of traditional 2D image data.

4.5.1 Trade-offs and Model Suitability for Syllable Classification

Each model demonstrates unique strengths and weaknesses, making them suitable for different applications:

ANN Model: Best suited for smaller datasets or applications requiring fast training and inference. Its strength lies in efficiency, though it may struggle with complex feature extraction.

LSTM Model: Excels in sequential classification tasks, making it a strong candidate for syllable-based voice recognition, where syllable order and dependencies matter.

CNN Model: Achieves the highest accuracy due to superior feature extraction, but requires longer training and higher computational resources, making it less ideal for real-time or low-power applications.

From these observations, it is evident that while CNN achieved the highest overall accuracy, it required significantly more training epochs compared to LSTM and ANN. In contrast, LSTM demonstrated a strong balance between accuracy and efficiency, making it an ideal candidate for syllable-based voice recognition, particularly when processing sequential audio data. Meanwhile, ANN remains a viable option for simpler tasks where rapid training is prioritized over sequential dependencies.

This comparative analysis highlights the importance of selecting a model based on the specific requirements of the task, considering factors such as accuracy, training time, computational cost, and model interpretability.

While CNN demonstrated the highest accuracy (99.68%), its computational complexity makes it less suitable for real-time applications. LSTM effectively captured temporal dependencies but required more training time than ANN. Additionally, the dataset's limited size may have led to overfitting, and future studies should explore transfer learning to mitigate this issue. Despite these limitations, the findings provide a foundation for future research in syllable-based recognition for low-resource languages.

5. CONCLUSION

This study evaluated the effectiveness of Artificial Neural Networks (ANN), Long Short-Term Memory Networks (LSTM), and Convolutional Neural Networks (CNN) for syllable-based voice recognition in Indonesian. Using a manually segmented dataset of 309 syllable parts, the models were compared across key performance metrics, including accuracy, and confusion metrics. In terms of dataset size, it consisted of 309 syllable parts, which, while sufficient for this study, may limit generalization to larger vocabularies or more complex datasets. Although accurate, manual segmentation is time-intensive and not scalable for larger datasets, posing a barrier for broader applications. For the class Imbalance, some syllable parts were underrepresented, impacting the accuracy of the prediction.

Challenges encountered during this research included the time-consuming process of manual labeling and the limited dataset size. Even though **the experiment faced** those challenges, the study successfully demonstrated the feasibility of syllable-based voice recognition using 1D array data from raw audio recordings. Based on the experiment that had been conducted, the CNN model is the best model to process the voice recording classification with 99.68% accuracy, followed by LSTM with 93.85% accuracy and ANN with 91.91% accuracy.

Furthermore, this research may contribute to the advancement in Voice Recognition for Low-Resource Languages and expand voice recognition research by focusing on Indonesian, a low-resource language, providing a foundation for future studies. The findings highlight the strengths and limitations of ANN, LSTM, and CNN for syllable recognition, offering guidance for selecting models based on specific requirements. The dataset created for this study can serve as a benchmark for future research on Indonesian syllable recognition or for developing automated segmentation tools.

REFERENCES

[1]. M. M. Abdulghani, W. L. Walters, and K. H. Abed, "Electroencephalography-Based Inner Speech Classification Using LSTM and Wavelet Scattering Transformation (WST)," in *Contemporary Perspective on Science, Technology and Research*, vol. 3, pp. 29–52, B P International, 2024, doi: 10.9734/bpi/cpstr/v3/6989c.

[2]. H. Abdullah, W. Garcia, C. Peeters, P. Traynor, K. R. B. Butler, and J. Wilson, "Practical Hidden Voice Attacks against Speech and Speaker Recognition Systems," unpublished.

[3]. Ç. Bakır, "Automatic Voice and Speech Recognition System for the German Language with Deep Learning Methods," Original Research Paper, Adv. Technol. Sci., no. 4, 2013. [Online]. Available: http://ijamec.atscience.org

[4]. O. Chandraumakantham, N. Gowtham, M. Zakariah, and A. Almazyad, "Multimodal Emotion Recognition Using Feature Fusion: An LLM-Based Approach," IEEE Access, vol. 12, pp. 108052–108071, 2024, doi: 10.1109/ACCESS.2024.3425953.

[5]. Z. Chen, J. Ge, H. Zhan, S. Huang, and D. Wang, "Pareto Self-Supervised Training for Few-Shot Learning," *unpublished*.

[6]. J. Dai, "Sparse Discrete Wavelet Decomposition and Filter Bank Techniques for Speech Recognition," 2019.

[7]. R. Ganchev and M. Informatics, "Voice Signal Processing for Machine Learning: The Case of Speaker Isolation Overview and Evaluation of Decomposition Methods Applied to the Input Signal of Voice Processing ML Models," unpublished.

[8]. B. Guan, J. Cao, X. Wang, Z. Wang, M. Sui, and Z. Wang, "Integrated Method of Deep Learning and Large Language Model in Speech Recognition," 2024. [Online]. Available: https://doi.org/10.20944/preprints202407.1520.v3

[9]. M. D. Hassan, A. Nejdet Nasret, M. R. Baker, and S. Mahmood, "Enhancement automatic speech recognition by deep neural networks," *Original Research*, vol. 9, no. 4, pp. 921–927, 2021.

[10]. H. Isyanto, A. Setyo Arifin, and M. Suryanegara, "Voice Biometrics for Indonesian Language Users using Algorithm of Deep Learning CNN Residual and Hybrid of DWT-MFCC Extraction Features," *Int. J. Adv. Comput. Sci. Appl. (IJACSA)*, vol. 13, no. 5. [Online]. Available: www.ijacsa.thesai.org

[11]. H. Y. Khdier, W. M. Jasim, and S. A. Aliesawi, "Deep Learning Algorithms based Voiceprint Recognition System in Noisy Environment," *J. Phys.: Conf. Ser.*, vol. 1804, no. 1, 2021.

[12]. S. Kumar Nayak, A. Kumar Nayak, S. Mishra, and P. Mohanty, "Deep Learning Approaches for Speech Command Recognition in a Low Resource KUI Language," *Int. J. Intell. Syst. Appl. Eng. (IJISAE)*, vol. 2023, no. 2. [Online]. Available: www.ijisae.org

[13]. K. Li, A. Zhu, Song, Zhao, Liu, and Jiabei, "Utilizing Deep Learning to Optimize Software Development Processes," *J. Comput. Technol. Appl. Math.*, vol. 1, no. 1, 2024, doi: 10.5281/zenodo.11084103.

[14]. K.-Y. Liu, S.-S. Wang, Y. Tsao, and J.-W. Hung, "Speech enhancement based on the integration of fully convolutional network, temporal lowpass filtering and spectrogram masking," *unpublished*.

[15]. Z. Ma *et al.*, "An Embarrassingly Simple Approach for LLM with Strong ASR Capacity," *unpublished*.

[16]. V. Mitra *et al.*, "Robust Features in Deep Learning-Based Speech Recognition," *unpublished*.

[17]. A. Moondra and P. Chahal, "Improved Speaker Recognition for Degraded Human Voice using Modified-MFCC and LPC with CNN," *Int. J. Adv. Comput. Sci. Appl. (IJACSA)*, vol. 14, no. 4. [Online]. Available: www.ijacsa.thesai.org

[18]. M. Novela and T. Basaruddin, "Dataset Suara dan Teks Berbahasa Indonesia pada Rekaman Podcast dan Talk Show," *Agustus*, vol. 11, no. 2, pp. 61–66.

[19]. T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, 2003, doi: 10.1016/S0167-6393(03)00099-2.

[20]. N. N. Prachi, F. M. Nahiyan, M. Habibullah, and R. Khan, "Deep Learning Based Speaker Recognition System with CNN and LSTM Techniques," in *Proc. 2022 Int. Conf. Interdiscip. Res. Technol. Manag. (IRTM)*, 2022, doi: 10.1109/IRTM54583.2022.9791766.

[21]. F. M. Rammo and M. N. Al-Hamdani, "Detecting the Speaker Language Using CNN Deep Learning Algorithm," *Iraqi J. Comput.*

Sci. Math., vol. 3, no. 1, pp. 43–52, 2022, doi: 10.52866/ijcsm.2022.01.01.005.

[22]. M. Ravanelli and Y. Bengio, "Speaker Recognition from Raw Waveform with SincNet." [Online]. Available: https://github.com/mravanelli/SincNet/

[23]. I. Rebai, Y. Benayed, W. Mahdi, and J. P. Lorré, "Improving speech recognition using data augmentation and acoustic model fusion," *Procedia Comput. Sci.*, vol. 112, pp. 316–322, 2017, doi: 10.1016/j.procs.2017.08.003.

[24]. T. N. Sainath *et al.*, "Multichannel Signal Processing with Deep Neural Networks for Automatic Speech Recognition," *unpublished*.

[25]. K. A. Salman, K. Shaker, and J. J. Stephan, "Speaker Recognition Using Deep Neural Networks with Combine Feature Extraction Techniques," *Periodicals of Engineering and Natural Sciences*, vol. 7, no. 2, pp. 10–13, 2019. [Online]. Available: http://pen.ius.edu.ba

[26]. S. Suyanto, A. Romadhony, F. Sthevanie, and R. N. Ismail, "Augmented words to improve a deep learning-based Indonesian syllabification," *Heliyon*, vol. 7, no. 10, 2021, doi: 10.1016/j.heliyon.2021.e08115.

[27]. Z. Tan, T. Chen, Z. Zhang, and H. Liu, "Sparsity-Guided Holistic Explanation for LLMs with Interpretable Inference-Time Intervention," 2024. [Online]. Available: www.aaai.org

[28]. S. Team, "Seed-ASR: Understanding Diverse Speech and Contexts with LLM-based Speech Recognition," *unpublished*.

[29]. V. Tiwari, "MFCC and its Applications in Speaker Recognition," *unpublished*.

[30]. L. T. Van, T. T. Le Dao, T. Le Xuan, and E. Castelli, "Emotional Speech Recognition Using Deep Neural Networks," *Sensors*, vol. 22, no. 4, 2022, doi: 10.3390/s22041414.

[31]. Y. Weng and J. Wu, "Big Data and Machine Learning in Defence," *Int. J. Comput. Sci. Inf. Technol.*, vol. 16, no. 2, pp. 25–35, 2024, doi: 10.5121/ijcsit.2024.16203.

[32]. G. Yang, Z. Ma, F. Yu, Z. Gao, S. Zhang, X. Chen, and M. Key, "MaLa-ASR: Multimedia-Assisted LLM-Based ASR." [Online]. Available: https://github.com/X-

[33]. S. Yang, Y. Zhao, and H. Gao, "Using Large Language Models in Real Estate Transactions: A Few-shot Learning Approach," *unpublished*.

[34]. D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*, Signals and Communication Technology Series. [Online]. Available: http://www.springer.com/series/4748

[35]. X. Zhang, R. Chowdhury, R. K. Gupta, and J. Shang, "Large Language Models for Time Series: A Survey." [Online]. Available: https://github.com/xiyuanzh/awesome-llm-time-series

[36]. Y. Zhao, S. Yang, and H. Gao, "Utilizing Large Language Models to Analyze Common Law Contract Formation," *unpublished*.

[37]. T. Wan Sen, K. H. Dewantara, C. Baru, and F. Komputer, "Data suara ucapan vokal Bahasa Indonesia," Information System Application, vol. 1, no. 2, n.d.

[38]. T. Wan and S. #1, "Frekuensi dominan dalam vokal Bahasa Indonesia," IT for Society, vol. 1, no. 2, n.d.