Comparative Analysis of MLP, CNN, and RNN Models in Automatic Speech Recognition: Dissecting Performance Metric

Abraham K. S. Lenson, Gregorius Airlangga
Information System Study Program, Engineering Faculty, Universitas Katolik Indonesia Atma Jaya, Indonesia

ARTICLE INFORMATION

Article History:

Submitted 25 November 2023
Revised 28 December 2023
Accepted 31 December 2023

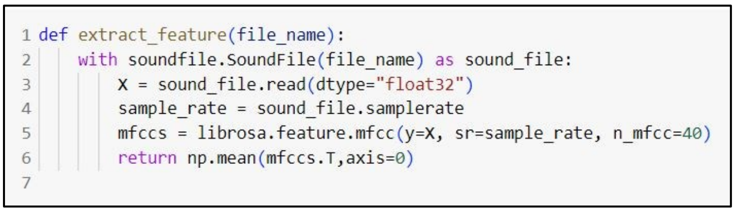
Keywords:

Machine Learning;
Deep Learning;
Speaker Identification;
MLP;
CNN;
RNN

Corresponding Author:

Gregorius Airlangga
Universitas Katolik Indonesia
Atma Jaya, Indonesia.
Email:
gregorius.airlangga@atmajaya.ac.id

ABSTRACT



```
def extract_feature(file_name):
    with soundfile.SoundFile(file_name) as sound_file:
        X = sound_file.read(dtype="float32")
        sample_rate = sound_file.samplerate
        mfccs = librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40)
        return np.mean(mfccs.T,axis=0)
```

This study conducts a comparative analysis of three prominent machine learning models: Multi-Layer Perceptrons (MLP), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) in the field of automatic speech recognition (ASR). This research is distinct in its use of the LibriSpeech 'test-clean' dataset, selected for its diversity in speaker accents and varied recording conditions, establishing it as a robust benchmark for ASR performance evaluation. Our approach involved preprocessing the audio data to ensure consistency and extracting Mel-Frequency Cepstral Coefficients (MFCCs) as the primary features, crucial for capturing the nuances of human speech. The models were meticulously configured with specific architectural details and hyperparameters. The MLP and CNN models were designed to maximize their pattern recognition capabilities, while the RNN (LSTM) was optimized for processing temporal data. To assess their performance, we employed metrics such as precision, recall, and F1-score. The MLP and CNN models demonstrated exceptional accuracy, with scores of 0.98 across these metrics, indicating their effectiveness in feature extraction and pattern recognition. In contrast, the LSTM variant of RNN showed lower efficacy, with scores below 0.60, highlighting the challenges in handling sequential speech data. The results of this study shed light on the differing capabilities of these models in ASR. While the high accuracy of MLP and CNN suggests potential overfitting, the underperformance of LSTM underscores the necessity for further refinement in sequential data processing. This research contributes to the understanding of various machine learning approaches in ASR and paves the way for future investigations. We propose exploring hybrid model architectures and enhancing feature extraction methods to develop more sophisticated, real-world ASR systems. Additionally, our findings underscore the importance of considering model-specific strengths and limitations in ASR applications, guiding the direction of future research in this rapidly evolving field.

Document Citation:

A. K. S. Lenson, G. Airlangga, "Comparative Analysis of MLP, CNN, and RNN Models in Automatic Speech Recognition: Dissecting Performance Metric," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 5, no. 4, pp. 576-583, 2023, DOI: 10.12928/biste.v5i4.9668.

1. INTRODUCTION

In the rapidly evolving landscape of artificial intelligence and digital signal processing, automatic speech recognition (ASR) has emerged as a pivotal field, increasingly vital in the realms of security, personalization of services, and human-computer interaction [1]–[3]. This domain, blending the intricacies of acoustic engineering and machine learning, is instrumental in a variety of cutting-edge applications, ranging from voice-activated systems and forensic analysis to customizing user experiences in smart devices [4]–[6]. The fundamental challenge in speaker identification lies in the nuanced task of accurately distinguishing individual voices amidst a plethora of audio signals, a feat that involves complex signal processing and advanced pattern recognition techniques [7]–[9]. Addressing this challenge are sophisticated machine learning models, which have become central to achieving high levels of precision and efficiency in ASR. Among these, Multi-Layer Perceptrons (MLP), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) stand out due to their distinct methodologies and analytical capabilities in processing speech [10]–[12]. Our study uniquely contributes by not only comparing these models' performance but also investigating their adaptability in varied speech processing scenarios. We have chosen MLP, CNN, and RNN due to their distinct methodologies and analytical capabilities in processing speech. The selection is justified by the need to understand how these models, each with different strengths, perform under the complex demands of speaker identification.

The field of ASR has undergone substantial evolution, marked by a transition from classical machine learning techniques to more sophisticated deep learning models. Earlier approaches, such as Gaussian Mixture Models (GMM) and Support Vector Machines (SVM), laid the groundwork but often fell short in capturing the complex dynamics of human speech [13]–[15]. The integration of neural networks, particularly CNNs and RNNs, has revolutionized this space. CNNs, with their deep learning architecture, have excelled in extracting detailed features from audio spectrograms, enhancing the system's ability to function effectively in acoustically diverse environments. RNNs, especially those incorporating LSTM units, have demonstrated proficiency in handling the temporal and sequential aspects of speech data, critical for differentiating individual speakers. MLPs, while more basic compared to CNNs and RNNs, continue to hold relevance due to their straightforward structure and computational efficiency, particularly in less demanding speaker identification tasks [16]–[18]. Despite these technological advancements, challenges remain, including the need for expansive and diverse training datasets, ensuring robustness against varying background noises, and enhancing the generalizability of models to handle different accents, dialects, and speaking styles [19],[20].

This study is committed to a comprehensive evaluation and comparison of MLP, CNN, and RNN models in the realm of speaker identification. It aims to transcend traditional performance metrics, delving into a deeper analysis of the models' feature extraction strengths, learning dynamics, adaptability to acoustic variations, and their computational efficiency and scalability. A significant focus will be placed on understanding the specific scenarios and conditions where each model exhibits superior performance, and how they can be optimally employed in practical, real-world applications [21]–[23]. The remaining of research paper is organized into the following, in the section 2, we will provide a review of existing research in ASR, emphasizing the evolution of machine learning models in this field. It will highlight significant studies, identify gaps in current research, and set the theoretical foundation for the study. Subsequently, in section 3, we will detail the research methodology, including data collection, preprocessing, feature extraction techniques, and the configuration of each machine learning model (MLP, CNN, RNN). It will also elaborate on the performance evaluation metrics and validation processes used in the study.

Experimental results and in-depth analysis will be discussed in the section 4, here, the research findings will be presented, encompassing a thorough quantitative analysis and comparative evaluation of MLP, CNN, and RNN models. This section will not only focus on performance metrics but also provide a qualitative analysis of why certain models perform better under specific conditions. Then, we will offer an extensive discussion of the results, interpreting the performance of each model in the context of current technological challenges and their practical implications in speaker identification systems. The final section will synthesize the key findings, contributions, and insights of the study, reflecting on the broader implications for the field of speaker identification and machine learning. It will also offer closing thoughts on the future trajectory of research in this area.

2. LITERATURE SURVEY

Speaker identification has undergone a significant transformation over the years, evolving from basic statistical models to sophisticated machine learning algorithms. Initial approaches in the field relied on Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) [24],[25]. These models, foundational in the early stages of ASR, were adept at handling structured, controlled speech but often struggled with the complexity and variability inherent in natural, spontaneous speech. The limited ability of these models to adapt to varying speech patterns, background noises, and emotional tones in speech led researchers to explore more adaptable solutions. The introduction of neural network models marked a paradigm shift in speaker identification. Multi-Layer Perceptron (MLP), one of the earliest forms of neural networks, brought a new perspective to the field. MLPs, despite their relatively simple architecture, were effective in extracting relevant features from speech data, particularly in more controlled environments [26],[27]. However, their non-temporal nature meant that they were less effective in capturing the dynamics and nuances of speech over time, a critical aspect of speaker identification. This limitation was addressed with the advent of Convolutional Neural Networks (CNN). CNNs, renowned for their feature extraction capabilities in image processing, were adapted to analyze complex audio signals [28],[29]. Their ability to process and learn from spectrograms of speech allowed for a more detailed and nuanced analysis of audio data. Studies demonstrated the superiority of CNNs in identifying unique speech patterns, particularly in environments with variable acoustic conditions.

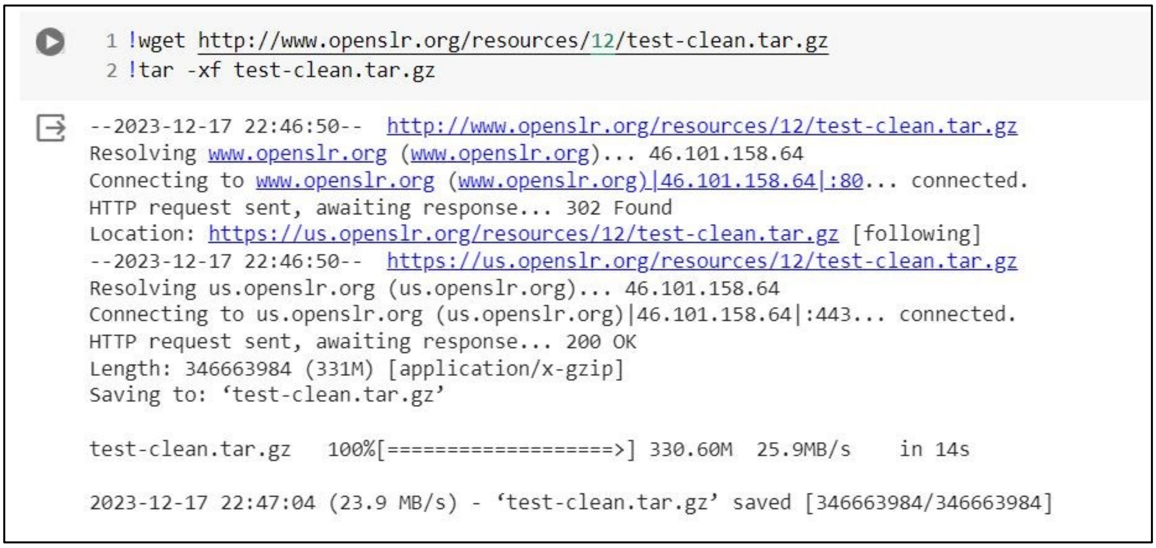
Recurrent Neural Networks (RNN), especially those with Long Short-Term Memory (LSTM) units, further advanced the field. RNNs' ability to process sequential and time-series data made them particularly suitable for speaker identification tasks [30]. LSTMs, capable of capturing long-term dependencies in data, addressed one of the significant challenges in speech processing, understanding the temporal dynamics and context within speech. This feature made RNNs and LSTMs particularly adept at differentiating between speakers based on the temporal patterns in their speech. The comparative analysis of MLP, CNN, and RNN models in speaker identification has been a crucial area of focus in recent research [31]–[33]. Studies have specifically looked at the strengths and weaknesses of each model type under various conditions. MLPs, while simpler and less resource-intensive, showed limitations in handling complex and variable speech data. CNNs, on the other hand, excelled in feature extraction but sometimes lacked in capturing long-term temporal dependencies. RNNs, particularly LSTMs, filled this gap but often required more computational resources.

The current state of the art in speaker identification is characterized by a push towards integrating and optimizing these various machine learning models to achieve higher accuracy, efficiency, and adaptability. Hybrid models that combine CNNs and RNNs are at the forefront, offering improved performance over traditional single-model approaches. However, challenges remain, particularly in ensuring the models' robustness across different languages, accents, and noisy environments. The need for large and diverse training datasets, the computational demands of more complex models, and the quest for real-time processing capabilities continue to drive research in this field. This review of the literature in speaker identification reveals a field that is in constant flux, with evolving models and techniques aimed at improving the accuracy and efficiency of speaker recognition systems. The comparative analysis of MLP, CNN, and RNN models, along with the exploration of proposed approaches, underscores the ongoing efforts to refine and optimize these technologies. As the field advances, the balance between model complexity, computational efficiency, and real-world applicability remains a key focus of research, setting the stage for future developments in speaker identification.

3. METHODOLOGY

3.1. Data Collection and Preprocessing

Our study utilizes the LibriSpeech "test-clean" dataset, a widely recognized benchmark in speaker identification research. This dataset was chosen for its comprehensive collection of labeled audio recordings, showcasing a diverse range of speaker accents and recording conditions, critical for testing the robustness of ASR models. Preprocessing was pivotal for ensuring data uniformity and optimizing model performance. We standardized each audio file's sample rate and format, preparing them for consistent and accurate feature extraction. The dataset was acquired using the following process shown in the Figure 1.



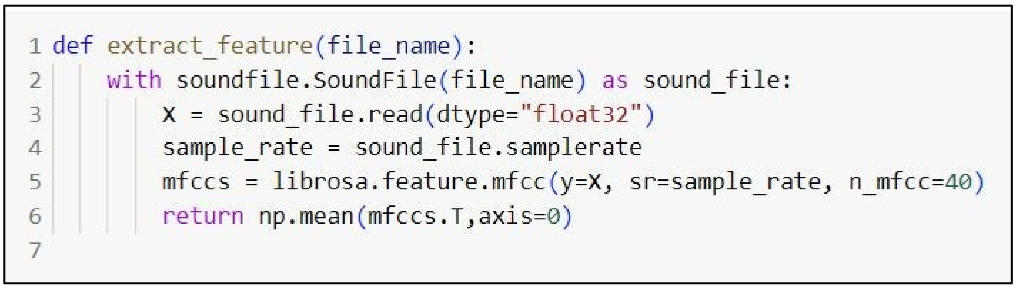
```
!wget http://www.openslr.org/resources/12/test-clean.tar.gz
!tar -xf test-clean.tar.gz
```

```
--2023-12-17 22:46:50--  http://www.openslr.org/resources/12/test-clean.tar.gz
Resolving www.openslr.org (www.openslr.org)... 46.101.158.64
Connecting to www.openslr.org (www.openslr.org)|46.101.158.64|:80... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://us.openslr.org/resources/12/test-clean.tar.gz [following]
--2023-12-17 22:46:50--  https://us.openslr.org/resources/12/test-clean.tar.gz
Resolving us.openslr.org (us.openslr.org)... 46.101.158.64
Connecting to us.openslr.org (us.openslr.org)|46.101.158.64|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 346663984 (331M) [application/x-gzip]
Saving to: ‘test-clean.tar.gz’

test-clean.tar.gz   100%[===================>] 330.60M  25.9MB/s    in 14s

2023-12-17 22:47:04 (23.9 MB/s) - ‘test-clean.tar.gz’ saved [346663984/346663984]
```

Figure 1. Data acquisition

3.2. Feature Extraction

Our study utilizes the LibriSpeech "test-clean" dataset, a widely recognized benchmark in speaker identification research. This dataset was chosen for its comprehensive collection of labeled audio recordings, showcasing a diverse range of speaker accents and recording conditions, critical for testing the robustness of ASR models. Preprocessing was pivotal for ensuring data uniformity and optimizing model performance. We standardized each audio file's sample rate and format, preparing them for consistent and accurate feature extraction.The extraction process is shown in the Figure 2.



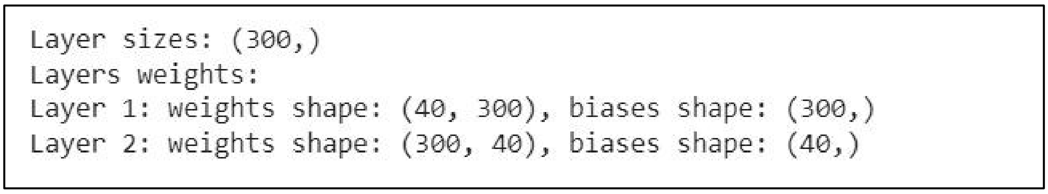
```
def extract_feature(file_name):
    with soundfile.SoundFile(file_name) as sound_file:
        X = sound_file.read(dtype="float32")
        sample_rate = sound_file.samplerate
        mfccs = librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40)
        return np.mean(mfccs.T,axis=0)

```

Figure 2. Feature extraction

3.3. Model Configuration

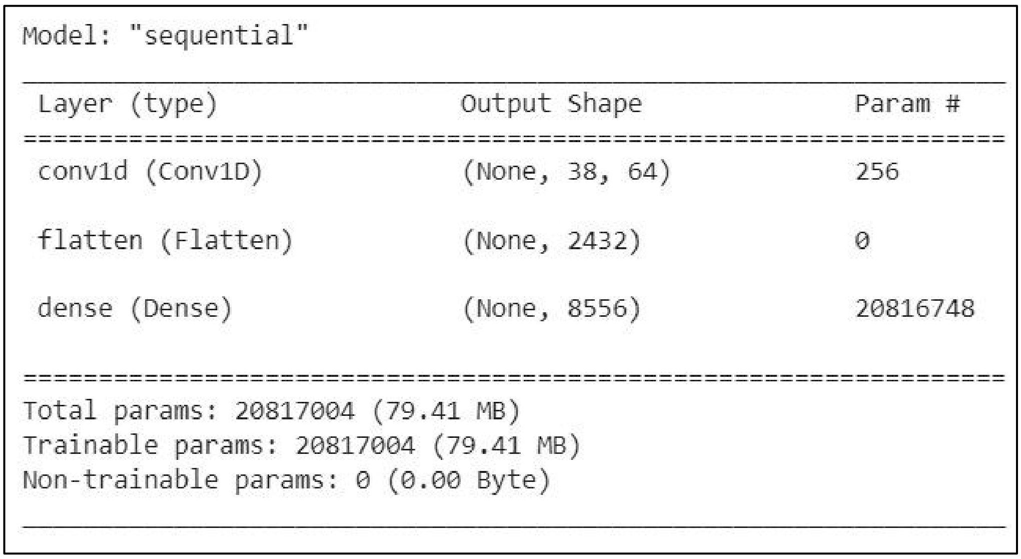
Three distinct models were configured, each representative of a different class of machine learning algorithms widely recognized in the speaker identification domain:

- MLP Classifier: The MLP was chosen for its simplicity and effectiveness in pattern recognition. It was configured with a multi-layer architecture, including two hidden layers, each with 128 neurons, and ReLU activation functions. This setup is designed to reduce the risk of overfitting while maintaining sufficient complexity for accurate speaker recognition is shown in the Figure 3.
- CNN: The CNN model was selected for its superior feature extraction capabilities, particularly adept at processing complex audio data. Our CNN architecture comprises convolutional layers with varying filter sizes to capture a range of speech signal features, followed by pooling layers and fully connected layers. This design aims to efficiently extract and learn representative features from the audio spectrograms is shown in the Figure 4.
- RNN (LSTM): The LSTM variant of RNN was chosen for its proficiency in processing sequential and temporal data, a key aspect of speech. Our LSTM model incorporates multiple layers with dropout regularization to enhance generalization. The architecture is tailored to track long-term dependencies in speech, a fundamental challenge in speaker identification is shown in the Figure 5.



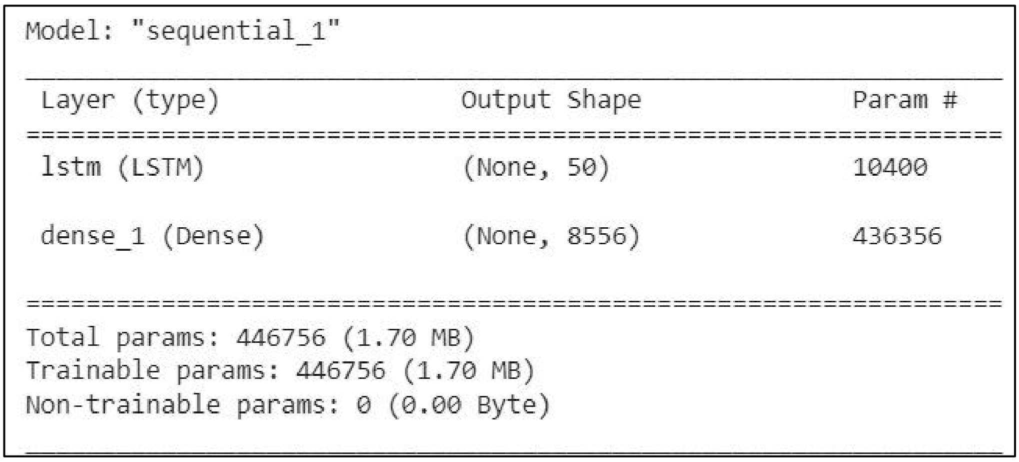
```
Layer sizes: (300,)
Layers weights:
Layer 1: weights shape: (40, 300), biases shape: (300,)
Layer 2: weights shape: (300, 40), biases shape: (40,)
```

Figure 3. MLP architecture



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 38, 64)            256

 flatten (Flatten)           (None, 2432)              0

 dense (Dense)               (None, 8556)              20816748

=================================================================
Total params: 20817004 (79.41 MB)
Trainable params: 20817004 (79.41 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 4. CNN architecture



```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 50)                10400

 dense_1 (Dense)             (None, 8556)              436356

=================================================================
Total params: 446756 (1.70 MB)
Trainable params: 446756 (1.70 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 5. LSTM architecture

3.4. Performance Evaluation Metrics and Validation

To evaluate model performance, we employed accuracy, precision, recall, and F1-score. These metrics were chosen for their comprehensiveness in assessing both the correctness and the completeness of the models' predictions. Additionally, confusion matrices were utilized to provide a detailed view of each model's performance across different speaker identities, offering insights into specific areas of strength and weakness.

3.5. Training and Testing

The dataset was split into training and testing sets, adhering to the standard 80/20 ratio. This split not only ensured a robust training phase but also facilitated an unbiased evaluation of the models' generalization capabilities on unseen data. The training process for each model was carefully monitored to prevent overfitting, a common pitfall in machine learning studies. The testing phase was conducted under strict controls to ensure the validity and reliability of the results. Cross-validation strategies were also employed to assess the models' generalizability and robustness across various subsets of the dataset.

4. RESULTS AND DISCUSSIONS

Our investigation into MLP, CNN, and RNN/LSTM models for speaker identification revealed marked differences in performance. As illustrated in Table 1, both MLP and CNN achieved high scores (precision, recall, and F1-score of 0.98), showcasing their proficiency in feature extraction and pattern recognition within complex audio datasets. These models demonstrated not only effective learning from training data but also strong generalization to new, unseen data, highlighting their potential for real-world applications. Conversely, the RNN (LSTM) model exhibited lower performance, with a precision of 0.58, recall of 0.56, and an F1-score of 0.54. This underperformance draws attention to specific challenges in the LSTM model's processing of sequential and temporal speech data. Potential limitations in the model's architecture, such as the number of LSTM layers or the configuration of hidden units, may have contributed to this discrepancy. Additionally, the LSTM's performance suggests a need for more refined feature engineering, particularly in capturing the temporal dynamics of speech, which are crucial for accurate speaker identification.

In addressing the LSTM's shortcomings, we propose exploring modifications in its architectural design, such as varying the depth and breadth of LSTM layers or employing different activation functions. Enhanced feature engineering, focusing on temporal aspects of speech, could also bolster LSTM's performance. Furthermore, the exploration of hybrid models that blend CNN's feature extraction prowess with RNN's temporal processing abilities is a promising direction. Such hybrid architectures could leverage the strengths of both model types, potentially leading to more sophisticated and accurate speaker identification systems. Future research should also consider the creation and utilization of more diverse datasets, encompassing a wider range of speech patterns, languages, and environmental conditions. This approach would enable a more comprehensive assessment of models' generalization and robustness, providing insights into their real-world applicability.

Our findings underscore the importance of continuous evaluation and optimization of ASR models to enhance their accuracy and practicality in various scenarios. While MLP and CNN have shown promising results, their deployment in real-world environments must account for factors such as varying noise levels and diverse accents. The study's reliance on the LibriSpeech "test-clean" dataset, though valuable, may not fully represent the complexity of real-world audio conditions. Therefore, further testing of these models under more challenging conditions is crucial to ensure their robustness and effectiveness in real-life applications. This study contributes significantly to understanding the capabilities and limitations of MLP, CNN, and RNN/LSTM models in speaker identification. The superior performance of MLP and CNN models illustrates their potential in this field, while the challenges faced by the LSTM model highlight the complexities of processing temporal speech characteristics. These insights are instrumental in guiding future advancements in ASR technologies, emphasizing the need for innovative approaches to address the evolving challenges in speaker identification.

Table 1. Results and Discussion

Model	Precision	Recall	F1-Score
MLP	0.98	0.98	0.98
CNN	0.98	0.98	0.98
RNN	0.58	0.56	0.54

5. CONCLUSIONS

In our study, we delved into the realm of speaker identification, a field of growing significance in the landscape of artificial intelligence and digital signal processing. Our focus was to conduct a rigorous comparative analysis of three distinct machine learning models: Multi-Layer-Perceptrons (MLP), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN), with a particular emphasis on Long Short-Term Memory (LSTM) networks. This exploration was rooted in the objective to understand and evaluate the intricacies and efficacies of these models in accurately identifying speakers. The results from our experiments presented a fascinating dichotomy in performance. Both the MLP and CNN models exhibited exceptional precision, recall, and F1-scores, each achieving near-perfect metrics of 0.98. This level of accuracy was indicative of their robust capabilities in handling the complex task of speaker identification, excelling in feature extraction and pattern recognition – critical components in processing and interpreting audio data. However, these remarkable scores also necessitated a prudent approach in their interpretation. The possibility of overfitting loomed, a scenario where models, despite their high accuracy on test data, might falter in generalizing to new, unseen datasets. In stark contrast, the LSTM model, a type of RNN designed specifically to handle sequential and temporal data, showed significantly lower performance. With precision, recall, and F1-scores noticeably below the 0.60 mark, the LSTM's results highlighted the nuanced and intricate nature of temporal data processing in speech. This pointed towards potential areas for model optimization and raised questions about the LSTM's ability to effectively capture and process the dynamic aspects of spoken language. Drawing from these findings, our study opens several avenues for future research and development in speaker identification. A critical area is the enhancement of feature engineering techniques. There is a clear opportunity to explore more advanced and sophisticated methods that could better encapsulate the complexities of speech, particularly for models like LSTM that specialize in processing sequential data. Another promising direction is the development of hybrid models that combine the strengths of CNNs in feature extraction with the sequential data processing capabilities of RNNs. Such hybrid models could potentially offer a more comprehensive and effective approach to speaker identification. The study also underscores the importance of focusing on model generalization and robustness. The high scores achieved by MLP and CNN models, while impressive, necessitate a careful approach to ensure these models can perform equally well in real-world scenarios, which often present more variability and complexity than controlled test environments. Furthermore, exploring alternative machine learning architectures, such as attention-based models or Transformers, presents fertile ground for innovation. These architectures, which have shown remarkable results in other areas of natural language processing, could offer new insights and improvements in the field of speaker identification. Finally, a key focus for future research should be the practical application and scalability of these models. Adapting and testing them in real-world scenarios, considering factors like computational efficiency, scalability, and adaptability to varying speech and noise conditions, will be vital in evolving these models from experimental frameworks to practical, deployable systems.

REFERENCES

[1] D. Sayers *et al.*, “The Dawn of the Human-Machine Era: A forecast of new and emerging language technologies,” vol. 1, no. hal-03230287, 2021, https://doi.org/10.17011/jyx/reports/20210518/1.

[2] Z. Lyu, “State-of-the-Art Human-Computer-Interaction in Metaverse,” *Int. J. Human--Computer Interact.*, pp. 1–19, 2023, https://doi.org/10.1080/10447318.2023.2248833.

[3] M. Schmitt and I. Flechais, “Digital Deception: Generative Artificial Intelligence in Social Engineering and Phishing,” *arXiv Prepr. arXiv2310.13715*, 2023, https://doi.org/10.2139/ssrn.4602790.

[4] M. Sadaf *et al.*, “Connected and Automated Vehicles: Infrastructure, Applications, Security, Critical Challenges, and Future Aspects,” *Technologies*, vol. 11, no. 5, p. 117, 2023, https://doi.org/10.3390/technologies11050117.

[5] K. Ashok and S. Gopikrishnan, “Statistical Analysis of Remote Health Monitoring Based IoT Security Models \& Deployments From a Pragmatic Perspective,” *IEEE Access*, vol. 11, pp. 2621–2651, 2023, https://doi.org/10.1109/ACCESS.2023.3234632.

[6] A. A. F. Alshadidi *et al.*, “Investigation on the Application of Artificial Intelligence in Prosthodontics,” *Appl. Sci.*, vol. 13, no. 8, p. 5004, 2023, https://doi.org/10.3390/app13085004.

[7] S. Ansari, K. A. Alnajjar, T. Khater, S. Mahmoud and A. Hussain, "A Robust Hybrid Neural Network Architecture for Blind Source Separation of Speech Signals Exploiting Deep Learning," in *IEEE Access*, vol. 11, pp. 100414-100437, 2023, https://doi.org/10.1109/ACCESS.2023.3313972.

[8] Van Hedger, S. C., Nusbaum, H. C., Heald, S. L., Huang, A., Kotabe, H. P., & Berman, M. G. (2019). The aesthetic preference for nature sounds depends on sound object recognition. *Cognitive science*, *43*(5), e12734, 2019, https://doi.org/10.1111/cogs.12734.

[9] A. Giachanou, P. Rosso, and F. Crestani, F. (2021). The impact of emotional signals on credibility assessment. *Journal of the Association for Information Science and Technology*, *72*(9), 1117-1132, 2021, https://doi.org/10.1002/asi.24480.

[10] O. I. Abiodun *et al.*, “Comprehensive review of artificial neural network applications to pattern recognition,” *IEEE access*, vol. 7, pp. 158820–158846, 2019, https://doi.org/10.1109/ACCESS.2019.2945545.

[11] N. Singh and H. Sabrol, “Convolutional neural networks-an extensive arena of deep learning. A comprehensive study,” *Arch. Comput. Methods Eng.*, vol. 28, no. 7, pp. 4755–4780, 2021, https://doi.org/10.1007/s11831-021-09551-4.

[12] M. M. Rahman *et al.*, “Prospective methodologies in hybrid renewable energy systems for energy prediction using artificial neural networks,” *Sustainability*, vol. 13, no. 4, p. 2393, 2021, https://doi.org/10.3390/su13042393.

[13] S. Sun, Z. Cao, H. Zhu and J. Zhao, "A Survey of Optimization Methods From a Machine Learning Perspective," in *IEEE Transactions on Cybernetics*, vol. 50, no. 8, pp. 3668-3681, 2020, https://doi.org/10.1109/TCYB.2019.2950779.

[14] J. Zhang, Z. Yin, P. Chen, and S. Nichele, “Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review,” *Inf. Fusion*, vol. 59, pp. 103–126, 2020, https://doi.org/10.1016/j.inffus.2020.01.011.

[15] C. Hema and F. P. G. Marquez, “Emotional speech recognition using cnn and deep learning techniques,” *Appl. Acoust.*, vol. 211, p. 109492, 2023, https://doi.org/10.1016/j.apacoust.2023.109492.

[16] Y. Qian, R. Ubale, P. Lange, K. Evanini, V. Ramanarayanan, and F. K. Soong, “Spoken language understanding of human-machine conversations for language learning applications,” *J. Signal Process. Syst.*, vol. 92, pp. 805–817, 2020, https://doi.org/10.1007/s11265-019-01484-3.

[17] S. P. Yadav, S. Zaidi, A. Mishra, and V. Yadav, "Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN)," *Arch. Comput. Methods Eng.*, vol. 29, no. 3, pp. 1753–1770, 2022, https://doi.org/10.1007/s11831-021-09647-x.

[18] Y. Lin, D. Guo, J. Zhang, Z. Chen, and B. Yang, "A unified framework for multilingual speech recognition in air traffic control systems," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 8, pp. 3608–3620, 2020, https://doi.org/10.1109/TNNLS.2020.3015830.

[19] A. Pervaiz *et al.*, "Incorporating noise robustness in speech command recognition by noise augmentation of training data," *Sensors*, vol. 20, no. 8, p. 2326, 2020, https://doi.org/10.3390/s20082326.

[20] Z. Almutairi and H. Elgibreen, "A review of modern audio deepfake detection methods: challenges and future directions," *Algorithms*, vol. 15, no. 5, p. 155, 2022, https://doi.org/10.3390/a15050155.

[21] G. Dulac-Arnold, N. Levine, D. J. Mankowitz, J. Li, C. Paduraru, S. Gowal, and T. Hester, "Challenges of real-world reinforcement learning: definitions, benchmarks and analysis," *Machine Learning*, vol. 110, no. 9, pp. 2419-2468, 2021, https://doi.org/10.1007/s10994-021-05961-4.

[22] Z. Xi *et al.*, "The rise and potential of large language model based agents: A survey," *arXiv Prepr. arXiv2309.07864*, 2023, https://doi.org/10.48550/arXiv.2309.07864.

[23] H. Du *et al.*, "Beyond deep reinforcement learning: A tutorial on generative diffusion models in network optimization," *arXiv Prepr. arXiv2308.05384*, 2023, https://doi.org/10.48550/arXiv.2308.05384.

[24] A. M. Deshmukh, "Comparison of hidden markov model and recurrent neural network in automatic speech recognition," *Eur. J. Eng. Technol. Res.*, vol. 5, no. 8, pp. 958–965, 2020, https://doi.org/10.24018/ejeng.2020.5.8.2077.

[25] S. Mao, D. Tao, G. Zhang, P. C. Ching, and T. Lee, "Revisiting hidden Markov models for speech emotion recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6715–6719, 2019, https://doi.org/10.1109/ICASSP.2019.8683172.

[26] S. Adams and P. A. Beling, "A survey of feature selection methods for Gaussian mixture models and hidden Markov models," *Artif. Intell. Rev.*, vol. 52, pp. 1739–1779, 2019, https://doi.org/10.1007/s10462-017-9581-3.

[27] Mustaqeem and S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, p. 183, 2019, https://doi.org/10.3390/s20010183.

[28] Z. Tariq, S. K. Shah, and Y. Lee, "Feature-based fusion using CNN for lung and heart sound classification," *Sensors*, vol. 22, no. 4, p. 1521, 2022, https://doi.org/10.3390/s22041521.

[29] F. Demir, D. A. Abdullah, and A. Sengur, "A new deep CNN model for environmental sound classification," *IEEE Access*, vol. 8, pp. 66529–66537, 2020, https://doi.org/10.1109/ACCESS.2020.2984903.

[30] S. Das, A. Tariq, T. Santos, S. S. Kantareddy, and I. Banerjee, "Recurrent Neural Networks (RNNs): Architectures, Training Tricks, and Introduction to Influential Research," *Mach. Learn. Brain Disord.*, pp. 117–138, 2023, https://doi.org/10.1007/978-1-0716-3195-9_4.

[31] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Inf. Fusion*, p. 101869, 2023, https://doi.org/10.1016/j.inffus.2023.101869.

[32] S. A. Syed, M. Rashid, S. Hussain, and H. Zahid, "Comparative analysis of CNN and RNN for voice pathology detection," *Biomed Res. Int.*, vol. 2021, pp. 1–8, 2021, https://doi.org/10.1155/2021/6635964.

[33] I. A. Thukroo, R. Bashir, and K. J. Giri, "A review into deep learning techniques for spoken language identification," *Multimed. Tools Appl.*, vol. 81, no. 22, pp. 32593–32624, 2022, https://doi.org/10.1007/s11042-022-13054-0.

AUTHOR BIOGRAPHY

GREGORIUS AIRLANGGA
Received the B.S. degree in information system from the Yos Sudarso Higher School of Computer Science, Purwokerto, Indonesia, in 2014, and the M.Eng. degree in informatics from Atma Jaya Yogyakarta University, Yogyakarta, Indonesia, in 2016. He got Ph.D. degree with the Department of Electrical Engineering, National Chung Cheng University, Taiwan. He is also an Assistant Professor with the Department of Information System, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia. His research interests include artificial intelligence and software engineering include path planning, machine learning, natural language processing, deep learning, software requirements, software design pattern and software architecture.

ABRAHAM K. S. LENSON
Studied Information System in Atma Jaya Catholic University of Indonesia. In 2022, he earned a national scholarship and enrolled as an exchange student in University of Galway, Ireland. His technical interests includes software development, Linux system management, automation, and reverse engineering.