Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 2699~2709
e-ISSN: 2548-6861

Hyperparameter Optimization of CNN Based Open Set Speaker Verification Using MFCC and Speaker Embedding for Voice Biometric Security

Mirza Ardiana [1]*, Mat Syai'in [2], Alief Nur Aisyi Maulidhia [3]***, Aulia Rahma Annisa [4]**, Yudi Andika [5]****, Sholahuddin Muhammad Irsyad [6]***, Fauzan Izzul Haq [7]****

* Manajemen Bisnis, Jurusan Teknik Bangunan Kapal, Politeknik Perkapalan Negeri Surabaya
** Teknik Otomasi, Jurusan Teknik Kelistrikan Kapal, Politeknik Perkapalan Negeri Surabaya
*** Teknik Kelistrikan Kapal, Jurusan Teknik Kelistrikan Kapal, Politeknik Perkapalan Negeri Surabaya
**** Teknik Pengelasan, Jurusan Teknik Bangunan Kapal, Politeknik Perkapalan Negeri Surabaya
mirzaardiana@ppns.ac.id [1], matt.syaiin@ppns.ac.id [2], aliefnur@ppns.ac.id [3], auliaannisa@ppns.ac.id [4], yudi.andika@ppns.ac.id [5], mugammad.irsyad@ppns.ac.id [6], fauzanizzul13@student.ppns.ac.id [7]

Article Info

Article history:

Received 2026-05-18
Revised 2026-05-29
Accepted 2026-06-11

Keyword:

Audio Augmentation,
CNN,
Hyperparameter Optimization,
MFCC,
Speaker Embedding,
Voice Biometrics.

ABSTRACT

The development of voice based biometric security systems has increased the demand for authentication methods capable of operating accurately and securely in open set speaker verification scenarios. In this scenario, the system is required not only to recognize registered users but also to reject unknown users who are not included in the system database. This study focuses on hyperparameter optimization in a Convolutional Neural Network Embedding based speaker verification system using Mel Frequency Cepstral Coefficient (MFCC) features and speaker embeddings. The optimization process was conducted through several experimental stages, including MFCC parameter tuning, CNN architecture tuning, embedding dimension tuning, and audio augmentation analysis. The dataset consisted of Indonesian speech recordings from 8 registered speakers and 1 unknown speaker, sampled at 16 kHz under controlled recording conditions. The dataset was divided into training, enrollment, and testing subsets to support open set speaker verification evaluation and reduce data leakage. System performance was evaluated using accuracy, validation loss, False Acceptance Rate (FAR), False Rejection Rate (FRR), best threshold, and inference time. The experimental results show that the best configuration was achieved using the MFCC-C parameters (N_MFCC = 40, N_FFT = 1024, HOP_LENGTH = 256, N_MELS = 40), the CNN-E architecture with three convolution blocks (32-64-128), an embedding dimension of 64, and lightweight augmentation consisting of noise injection, pitch shifting, and time stretching. This configuration achieved stable system performance with a test accuracy of 96.43% and a FAR of 8.7%, while maintaining lightweight computational complexity and real time inference capability. The results also indicate that excessive augmentation may increase embedding overlap between speakers, thereby reducing system security performance. However, the study was conducted on a limited scale dataset and has not yet evaluated robustness against spoofing attacks, replay attacks, or adversarial synthesized voice attacks. Overall, the study indicates that hyperparameter optimization influences the balance between accuracy, computational efficiency, and biometric security performance in lightweight CNN based voice biometric authentication systems under limited scale evaluation conditions.

I. INTRODUCTION

The rapid development of digital security technology has increased the demand for authentication systems capable of providing higher security levels compared to conventional password and card based methods. One of the rapidly growing approaches is biometric technology, which identifies individuals based on unique characteristics such as facial features, fingerprints, iris patterns, and voice characteristics [1], [2], [3]. Among various biometric approaches has become an attractive solution because the authentication process can be performed naturally, without physical contact, and can be relatively easily implemented on devices with medium to low computational specifications.

Voice biometric technology has been widely applied in various fields, including room access systems [4], Internet of Things (IoT) devices [5], [6], smart home services [7], and artificial intelligence based security systems [8]. In such systems, authentication is performed by recognizing the user's voice characteristics through feature extraction and pattern matching processes against reference data stored in the system database. The success of the authentication process is highly dependent on the quality of voice feature representation and the model's ability to distinguish characteristics among different users.

In the field of speech signal processing, Mel Frequency Cepstral Coefficient (MFCC) is one of the most widely used feature extraction methods because it is capable of representing the spectral characteristics of human speech effectively [9], [10], [11]. MFCC features are subsequently processed using deep learning methods, particularly Convolutional Neural Networks (CNN), to generate higher level and more discriminative feature representations. The combination of MFCC and CNN has demonstrated promising performance in various speaker recognition and speaker verification studies [12], [13].

Nevertheless, speaker verification systems still face several challenges, particularly in open set speaker verification scenarios. In this condition, the system is not only required to recognize registered users but also to reject unknown speakers who are not included in the database. This issue becomes critical in security system implementations because falsely accepting unknown users may lead to unauthorized access to protected systems.

Most previous studies primarily focused on improving accuracy as the main performance indicator. However, in biometric authentication systems, security evaluation cannot rely solely on accuracy but must also consider security related metrics such as False Acceptance Rate (FAR) and False Rejection Rate (FRR). FAR represents the rate at which the system incorrectly accepts unauthorized users as legitimate users, whereas FRR indicates the rate at which the system incorrectly rejects legitimate users. These two parameters are important indicators for determining the security level and reliability of speaker verification systems.

In addition, the performance of speaker verification systems is strongly influenced by the hyperparameter configuration used. Parameters such as the number of MFCC coefficients, Fast Fourier Transform (FFT) size, hop length, number of CNN filters, number of convolution blocks, embedding dimension, and audio augmentation techniques directly affect feature extraction quality, model generalization capability, and authentication stability [14], [15], [16], [17]. However, comprehensive studies investigating hyperparameter optimization in open set speaker verification systems remain relatively limited.

Recent speaker verification systems have utilized advanced embedding architectures such as x vector, ECAPA-TDNN, wav2vec, and transformer based models. However, these approaches typically require significantly larger datasets and higher computational resources. This study focuses on lightweight MFCC based CNN embedding suitable for limited resource environments and real time authentication systems.

Therefore, this study proposes a lightweight CNN Embedding based speaker verification framework with hyperparameter optimization to improve verification performance while maintaining computational efficiency under limited scale open set biometric authentication scenarios.

Based on these conditions, this study focuses on hyperparameter optimization in a CNN based open set speaker verification system using MFCC features and speaker embeddings. The optimization process was conducted through several experimental stages, including MFCC parameter tuning, CNN architecture tuning, embedding dimension tuning, and audio augmentation analysis. System evaluation was performed using several performance metrics, including accuracy, validation loss, FAR, FRR, best threshold, and inference time, in order to obtain a model configuration that not only achieves good classification performance but also satisfies security requirements for real time voice authentication systems.

II. METHOD

This research proposes a hyperparameter optimization framework for an open set speaker verification system based on a Convolutional Neural Network (CNN) using Mel Frequency Cepstral Coefficient (MFCC) features and speaker embeddings as user voice identity representations. The proposed method is designed to obtain the best model configuration that not only achieves high classification performance but also satisfies the security requirements of a biometric voice authentication system.

In general, the research stages consist of dataset acquisition, audio preprocessing, dataset splitting, MFCC feature extraction, hyperparameter optimization, audio augmentation analysis, CNN model training, speaker embedding generation, open set speaker verification, and security evaluation using False Acceptance Rate (FAR) and False Rejection Rate (FRR). The overall research framework used in this study is illustrated in Figure 1.

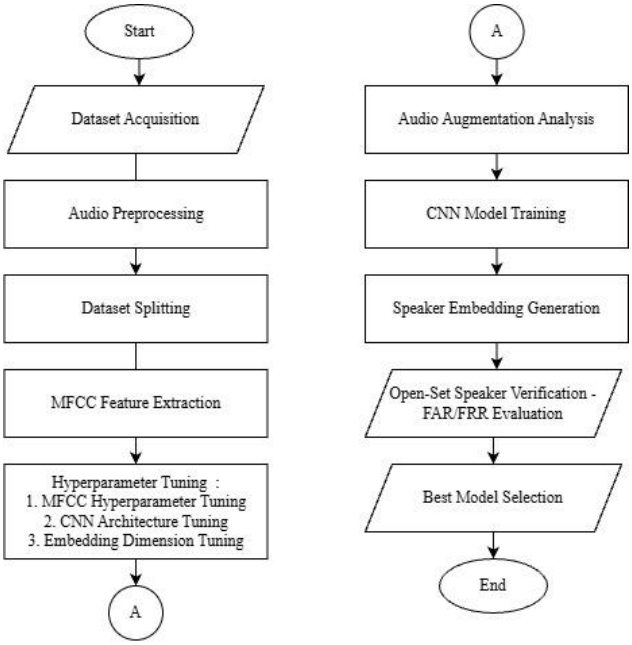

Figure 1. Research Flow

A. Dataset Acquisition

The dataset used in this study was obtained from Kaggle and consists of Indonesian language speech recordings with a sampling rate of 16,000 Hz and varying audio durations for each WAV recording, collected under relatively controlled recording conditions with limited environmental noise variations; audio preprocessing included fixed length standardization into 3 second segments, MFCC feature extraction, and feature normalization, although the dataset still contains limited speaker diversity and controlled recording conditions, which may limit the generalization capability of the proposed model in large scale real world speaker verification scenarios.

In this study, the dataset is divided into two main categories: registered speakers and unknown speakers. The registered speaker data consist of 8 speaker labels, namely LABEL0 to LABEL7, which are used as the primary identities during the training and testing processes. In addition, the study also includes one additional label, namely LABELUNKNOWN, which represents unknown users in the open set speaker verification scenario. Each label contains 23 audio samples (.wav files), resulting in a balanced data distribution across all speakers. Although the proposed system was evaluated using an open set scenario, the number of unknown speakers remains limited. Future work will involve larger and more diverse unknown speaker populations to better simulate real world authentication environments.

The inclusion of unknown speaker data aims to evaluate the system's ability to reject access attempts from users who are not registered in the speaker database. This approach is important in biometric security systems because, in real world conditions, the system is not only required to recognize authorized users but also expected to accurately detect and reject unauthorized or foreign users.

The use of voice biometrics in authentication systems is based on the unique characteristics of human voices, which are relatively difficult to replicate and allow authentication to be performed naturally without direct physical contact [18].

B. Audio Preprocessing

The preprocessing stage is performed to standardize the characteristics of the audio signals before the feature extraction process. All audio signals are converted into mono audio and resampled to 16 kHz to ensure consistent dataset specifications.

In addition, each audio signal is adjusted to a fixed duration of 3 seconds using trimming and zero padding processes. If the signal length is shorter than the target duration, zero values are added through padding. Conversely, if the signal length exceeds the target duration, the signal is truncated according to the predefined sample length.

This preprocessing stage is important to ensure consistent input dimensions for the CNN training process and to improve the stability of feature extraction [12].

C. Dataset Splitting

To avoid data leakage during the speaker verification process, the dataset is divided into three main subsets:

- *training set* 70%,
- *enrollment set* 15%,
- *testing set* 15%.

The training set is used for CNN model training so that the model can learn the characteristics of each speaker. The enrollment set is used to build the speaker database based on speaker embeddings, while the testing set is used to evaluate the system performance during classification and speaker verification.

Dataset splitting is performed using a stratified split method to maintain balanced data distribution across all speaker classes. This approach ensures consistent data representation during both training and evaluation stages.

Independent dataset separation is essential to prevent data leakage. Data leakage refers to a condition where information from the testing set or enrollment set is unintentionally used during the training process. This condition can cause the model to produce artificially high evaluation performance that does not accurately represent its real world capability. In other words, the model tends to "memorize" previously seen data, resulting in biased and unreliable evaluation results [19].

D. MFCC Feature Extraction

In this study, voice feature extraction is performed using the Mel Frequency Cepstral Coefficient (MFCC) method. MFCC is one of the most widely used feature extraction techniques in speech recognition and speaker verification systems because it effectively represents the spectral characteristics of human speech [13], [20].

The MFCC extraction process uses several main parameters, including:

- number of MFCC coefficients (N_MFCC),
- Fast Fourier Transform size (N_FFT),

- hop length,
- number of mel filters (N_MELS),
- Hann window function.

After extraction, the MFCC features are normalized using standard normalization to improve training stability. Several MFCC parameter tuning scenarios are evaluated in this study to analyze the effect of feature configuration on speaker verification performance.

E. Hyperparameter Optimization of MFCC Parameters

The first optimization stage focuses on MFCC feature extraction parameters. Several combinations of N_MFCC, N_FFT, HOP_LENGTH, and N_MELS are evaluated to analyze their influence on voice feature representation quality and speaker verification performance.

The selection of MFCC parameters significantly affects the spectral resolution and temporal representation of speech signals [21]. An optimal MFCC configuration is expected to produce more discriminative features, thereby improving the system's ability to distinguish speaker characteristics.

F. CNN Architecture Hyperparameter Optimization

The second optimization stage focuses on CNN architecture. This study evaluates several CNN configurations with variations in the number of convolution blocks, convolution filters, and dropout rates.

Each convolution block consists of a convolution layer, batch normalization, ReLU activation function, max pooling, and dropout layer. CNN is used because of its capability to learn spatial patterns from MFCC features and generate high level discriminative feature representations [22]. The evaluation is conducted to analyze the effect of network depth and model complexity on classification performance and speaker verification security.

G. Dimension Hyperparameter Optimization

The next optimization stage focuses on embedding dimensions. Embeddings are used to represent speaker identities as fixed dimensional numerical vectors before the softmax classification process.

Several embedding dimensions are evaluated to analyze their effect on inter speaker separation capability and verification stability. Larger embedding dimensions generally produce more detailed speaker representations; however, they also increase model complexity and the risk of overfitting [18], [23].

H. Audio Augmentation Analysis

To improve model generalization capability, this study applies several audio augmentation techniques to the training dataset. Augmentation is intended to simulate various real world environmental conditions that may affect voice quality during authentication. The augmentation techniques used include:

- noise injection,
- pitch shifting,
- time stretching,
- volume scaling,
- background noise.

Several augmentation combinations are evaluated to analyze their influence on the robustness of the open set speaker verification system.

Audio augmentation has been shown to improve the performance of deep learning models in speech processing systems by increasing the variability of training data distributions [24].

I. CNN Model Training

The CNN model is trained using MFCC features as the primary input. The CNN architecture consists of convolution layers followed by batch normalization and ReLU activation functions to extract high level features from speech signals.

At the final stage of the network, Global Average Pooling is employed to reduce parameter complexity and improve model generalization capability. The embedding output is then normalized using L2 normalization before being forwarded to the softmax classification layer.

The training process uses the Adam optimizer, sparse categorical cross entropy loss function, and early stopping based on validation loss.

The use of early stopping aims to reduce the risk of overfitting and preserve the best performing model during training [23].

J. Speaker Embedding Generation

After the CNN training process is completed, the speaker embedding model is constructed by extracting the output from the normalized embedding layer in the CNN architecture.

Speaker embedding vectors were extracted from the dense embedding layer of the CNN model after supervised training using speaker labels. The embedding layer generates compact latent feature representations that capture speaker specific characteristics from MFCC features. Before similarity evaluation, the embedding vectors were normalized using L2 normalization to improve cosine similarity stability. The dataset was divided into training, enrollment, and testing subsets to evaluate speaker verification performance under open set conditions. The enrollment subset was used to construct the speaker database, while the testing subset was used exclusively for verification evaluation to avoid data leakage.

This layer generates fixed dimensional numerical vector representations used as the distinctive identity of each speaker. In this study, the embedding dimension is determined through hyperparameter tuning to obtain the optimal speaker representation.

Mathematically, speaker embeddings are represented as:

$$e_i = [x_1, x_2, x_3, \dots, x_n] \quad (1)$$

where n denotes the embedding dimension and e_i represents the embedding vector of the i^{th} voice sample.

The speaker database is generated using the enrollment dataset. Each enrollment sample is converted into an embedding vector using the embedding model. Subsequently, embeddings belonging to the same speaker are averaged to

obtain the main speaker representation (speaker centroid). This process is formulated as:

$$e = \frac{1}{N}\sum_{i=1}^{N} e_i \qquad (2)$$

Where e represents the mean speaker embedding, N denotes the number of enrollment samples, and e_i represents each speaker embedding vector.

To improve matching stability, all embeddings are normalized using L2 normalization as follows:

$$\check{e} = \frac{e}{||e||_2} \qquad (3)$$

where,

$$||e||_2 = \sqrt{\sum_{i=1}^{N} e_i^2} \qquad (4)$$

The normalization process reduces the influence of embedding magnitude differences so that similarity measurements become more stable.

During speaker verification, cosine similarity is used to measure the similarity between the test embedding and the speaker embeddings stored in the database. The cosine similarity equation is expressed as follows [18]:

$$Similarity(A,B) = \frac{A.B}{||A||\,||B||} \qquad (5)$$

The resulting similarity value is then compared with a predefined security threshold to determine whether the speaker is classified as an authorized user or an unknown speaker. This approach enables the system to operate in an open set speaker verification scenario, where the system must recognize legitimate users while simultaneously rejecting unregistered users.

K. Open Set Speaker Verification and FAR/FRR Evaluation

The speaker verification system in this study employs an open set verification approach using cosine similarity between the test speaker embedding and the speaker database embeddings.

For each test sample, the system selects the highest similarity score as the verification result. If the similarity score exceeds a predefined threshold, the speaker is classified as an authorized user. Otherwise, the speaker is categorized as an unknown user.

The security performance of the system is evaluated using the following metrics:

- False Acceptance Rate (FAR)

$$FAR = \frac{FP}{FP+TN} \qquad (6)$$

- False Rejection Rate (FRR).

$$FRR = \frac{FN}{FN+TP} \qquad (7)$$

FAR indicates the rate at which the system incorrectly accepts unauthorized users as legitimate users, while FRR represents the rate at which the system incorrectly rejects legitimate users [25].

III. Results and Discussion

In this study, the experimental process was conducted progressively to obtain the optimal hyperparameter configuration for the CNN based open set speaker verification system using MFCC and speaker embedding. The evaluation focused on several key parameters, including MFCC tuning, CNN architecture tuning, embedding dimension tuning, and audio augmentation analysis.

A. MFCC Tuning

The first experiment in this study focused on tuning the Mel Frequency Cepstral Coefficient (MFCC) parameters to obtain the most optimal feature extraction configuration for the open set speaker verification system. Several MFCC configurations were evaluated, including:

- MFCC A : N_MFCC=13, N_FFT=512, HOP_LENGTH=128, and N_MELS=13
- MFCC B : N_MFCC=20, N_FFT=1024, HOP_LENGTH=256, and N_MELS=20
- MFCC C : N_MFCC=40, N_FFT=1024, HOP_LENGTH=256, and N_MELS=40
- MFCC D : N_MFCC=64, N_FFT=2048, HOP_LENGTH=512, and N_MELS=64

During the MFCC tuning process, the remaining parameters were fixed, where the CNN architecture consisted of three convolution blocks with filter configurations of 16–32–64 and a dropout rate of 0.3, while the embedding dimension was set to 64. This approach was applied to ensure that the performance variations were only influenced by the MFCC configuration. The results of MFCC hyperparameter tuning are presented in Table I.

TABLE I
MFCC HYPERPARAMETER TUNING

Experi-ment	Test Accura-cy	Best Thres-hold	Inferen-ce Time (ms/step)	FAR	FRR
MFCC-A	0.6071	0.9	78	0.5652	0.3214
MFCC-B	0.8214	0.9	262	0.2174	0.2857
MFCC-C	0.7143	0.85	268	0	0.0714
MFCC-D	0.8571	0.9	272	0.3478	0.3571

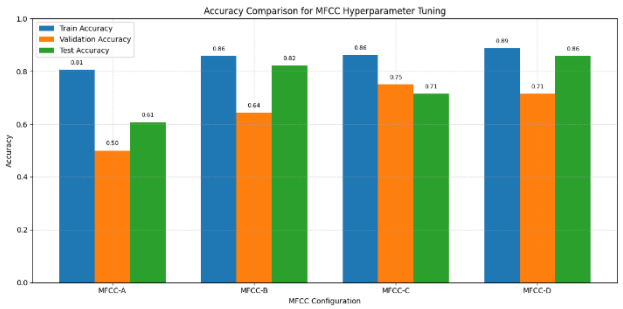

Figure 2. Accuracy Comparison for MFCC Hyperparameter Tuning

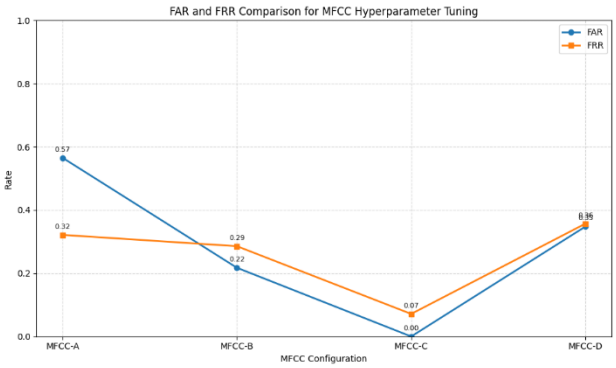

Figure 3. FAR and FRR for MFCC Hyperparameter Tuning

Based on the experimental results, the MFCC configuration significantly affected the speaker verification performance. MFCC-A produced the lowest performance due to the limited number of extracted features, while MFCC-B showed improvements in accuracy and a reduction in FAR. Although MFCC-D achieved the highest test accuracy, the FAR and FRR values increased considerably, indicating lower security performance in the open set verification scenario. Therefore, MFCC-C was selected as the optimal configuration because it achieved a FAR of 0% with a low FRR, resulting in better security performance and more stable generalization capability compared to the other configurations.

B. CNN Architecture Tuning

The second experiment focused on CNN architecture tuning to analyze the effect of convolution block depth, filter size, and dropout rate on the open set speaker verification performance. In this stage, the MFCC configuration used the best result from the previous experiment, namely MFCC-C with parameters N_MFCC = 40, N_FFT = 1024, HOP_LENGTH = 256, and N_MELS = 40, while the embedding dimension was fixed at 64 to ensure that performance variations were only caused by the CNN architecture.

The evaluated CNN architecture scenarios are described as follows:

- CNN-A = 2 Conv Blocks, Filters : 16-32, Dropout : 0.3
- CNN-B = 3 Conv Blocks, Filters : 16-32-64, Dropout : 0.3
- CNN-C = 4 Conv Blocks, Filters : 32-64-128-256, Dropout : 0.4
- CNN-D = 4 Conv Blocks, Filters : 16-32-64-128, Dropout : 0.35
- CNN-E = 3 Conv Blocks, Filters : 32-64-128, Dropout : 0.35

TABLE II
CNN ARCHITECTURE HYPERPARAMETER TUNING

Experi-ment	Test Accura-cy	Best Thres-hold	Inferen-ce Time (ms/step)	FAR	FRR
CNN-A	0.6071	0.95	198	0.4348	0.6071
CNN-B	0.7143	0.85	268	0	0.0714
CNN-C	1	0.9	408	0.2174	0.1071
CNN-D	1	0.9	263	0.1739	0.1071
CNN-E	0.9643	0.95	274	0.087	0.2857

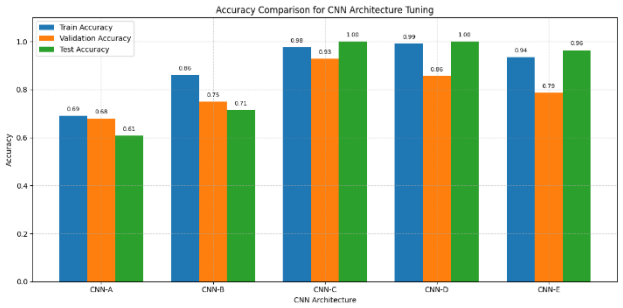

Figure 4. Accuracy Comparison for CNN Architecture Hyperparameter Tuning

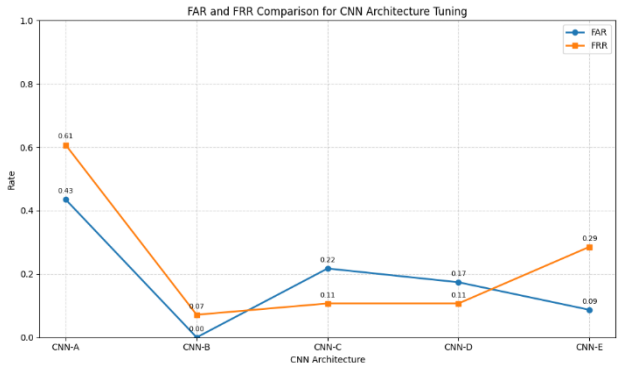

Figure 5. FAR and FRR for CNN Architecture Hyperparameter Tuning

The results indicate that each CNN architecture produced different performance characteristics. CNN-A achieved the lowest performance, indicating that a shallow network with limited filters could not adequately capture discriminative speaker representations. CNN-B improved the security performance by achieving a FAR of 0%, although its test accuracy remained relatively low.

CNN-C and CNN-D achieved the highest test accuracy of 100%. However, both configurations produced higher FAR values, indicating that deeper and more complex networks tended to overfit the training data, reducing the system's ability to reject unknown speakers. In addition, CNN-C had the highest inference time due to its deeper architecture and larger filter configuration.

The proposed CNN-Embedding (CNN-E) architecture consists of three convolution blocks with filter configurations

of 32, 64, and 128, respectively. Each convolution layer uses a kernel size of 3×3 with same padding and L2 regularization to improve feature generalization. Batch normalization and ReLU activation functions are applied after each convolution layer to stabilize the training process and improve nonlinear feature extraction capability. Max pooling layers with a pooling size of 2×2 are utilized to reduce feature dimensionality, followed by dropout regularization with a rate of 0.35 to minimize overfitting.

After the convolution blocks, a Global Average Pooling layer is used to generate compact feature representations before entering the embedding layer. The embedding layer consists of 64 dimensions with ReLU activation and L2 normalization before cosine similarity evaluation. The final classification layer uses softmax activation corresponding to the number of registered speakers. The model was trained using the Adam optimizer with a learning rate of 0.001 and sparse categorical cross entropy loss function.

CNN-E demonstrated the most balanced performance among all scenarios. This configuration achieved a test accuracy of 96.43% with a FAR of 8.7% and an inference time of 274 ms/step. The three convolution blocks with filters of 32–64–128 successfully generated more discriminative speaker representations without excessively increasing the model complexity.

Based on the overall results, CNN-E was selected as the best CNN architecture because it provided the best balance between classification performance, security performance, and computational efficiency. Therefore, CNN-E was chosen for the subsequent experiments and considered more suitable for real time biometric authentication systems.

C. Embedding Dimension Tuning

The third experiment analyzed the influence of embedding dimension size on the open set speaker verification performance. This stage used the best configuration obtained from the previous experiments, namely MFCC-C (40,1024,256,40) and CNN-E (32–64–128 with dropout 0.35). The remaining parameters were kept constant so that the performance variations were only influenced by the embedding dimension.

The evaluated embedding dimension scenarios are as follows:

- EMB-A = 32 Dimension
- EMB-B = 64 Dimension
- EMB-C = 128 Dimension
- EMB-D = 256 Dimension

Based on the experimental results presented in Table III, the EMB-A configuration with an embedding dimension of 32 produced the lowest performance among all evaluated scenarios. The model only achieved a test accuracy of 25% with a validation accuracy of 10.71%. In addition, the FAR reached 100%, indicating that all unknown speakers were incorrectly accepted as authorized users. This result suggests that a small embedding dimension could not generate sufficiently discriminative speaker representations, causing the model to fail in distinguishing speaker characteristics effectively.

TABLE III
EMBEDDING DIMENSION HYPERPARAMETER TUNING

Experiment	Test Accuracy	Best Threshold	Inference Time (ms/step)	FAR	FRR
EMB-A	0.25	0.5	256	1	0
EMB-B	0.9643	0.95	330	0.087	0.2857
EMB-C	0.8929	0.9	276	0.2609	0.3214
EMB-D	1	0.9	256	0.3478	0.0714

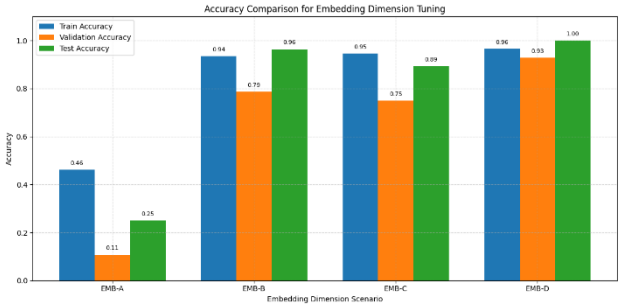

Figure 6. Accuracy Comparison for Embedding Dimension Hyperparameter Tuning

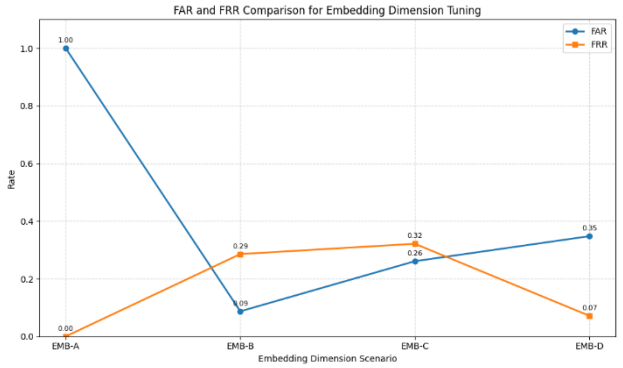

Figure 7. FAR and FRR Comparison for Embedding Dimension Hyperparameter Tuning

A significant performance improvement was observed in EMB-B with an embedding dimension of 64. This configuration achieved a test accuracy of 96.43% with a FAR of 8.7%. These results indicate that an embedding dimension of 64 was capable of generating sufficiently representative speaker embeddings without excessively increasing model complexity. Furthermore, the inference time of 330 ms/step remained relatively stable for real time implementation.

In the EMB-C configuration with an embedding dimension of 128, the model generated more detailed speaker representations. However, the security performance decreased, as indicated by the FAR of 26.09% and FRR of

32.14%. This finding demonstrates that increasing the embedding dimension does not always improve the model generalization capability. Instead, higher dimensional embeddings may increase overlap among speaker representations, resulting in less stable verification performance.

Meanwhile, EMB-D with an embedding dimension of 256 achieved the highest test accuracy of 100% and the lowest FRR of 7.14%. However, the FAR increased significantly to 34.78%, indicating that the system still frequently accepted unknown speakers as authorized users. In biometric authentication systems, a high FAR is considered a critical issue because it is directly related to the risk of unauthorized system access.

Based on the overall experimental results, EMB-B was selected as the optimal embedding dimension configuration. The selection was based on the balance between classification performance, security performance, and computational efficiency. Although EMB-D achieved higher accuracy, EMB-B produced a significantly lower FAR, making it more suitable for secure speaker verification based biometric authentication systems. In addition, the embedding dimension of 64 maintained a relatively lightweight model complexity, which is more appropriate for real time applications. The embedding dimension size of 64 was selected as a compromise between computational efficiency and discriminative capability. Lower embedding dimensions reduce memory usage and inference complexity, while excessively large embedding spaces may increase computational overhead and embedding overlap risks on limited-scale datasets.

D. Augmented Analysis

This stage analyzed the influence of various audio augmentation techniques on the performance of the open set speaker verification system. The experiments were conducted using the best hyperparameter configuration obtained from the previous stages, namely MFCC-C (40,1024,256,40), CNN-E (32–64–128 with dropout 0.35), and an embedding dimension of 64.

The initial augmentation stage employed three basic augmentation techniques, namely noise augmentation, pitch shifting, and time stretching. Subsequently, more detailed augmentation scenarios were developed, including small noise, medium noise, pitch up, pitch down, time stretch slow, time stretch fast, volume up, volume down, and background noise. These experiments were conducted to analyze the influence of augmentation type and quantity on the speaker authentication performance in the open set verification scenario.

Based on the experimental results presented in Table IV, the augmentation strategy significantly affected the speaker verification performance. Aug A achieved the most stable performance with a test accuracy of 96.43% and the lowest FAR of 8.7%. This result indicates that lightweight augmentation using limited combinations of noise, pitch shifting, and time stretching improved the model generalization capability without excessively altering the original speaker characteristics.

TABLE IV
AUGMENTED ANALYSIS

Code	Augmentation Type	Test Accuracy	Inferen-ce Time (ms/step)	FAR	FRR
Aug A	Noise, Pitch, Stretch	96.43%	290	8.70%	28.57%
Aug B	Small Noise, Medium Noise, Background Noise	89.29%	279	21.74%	35.71%
Aug C	Pitch Up, Pitch Down	10.71%	275	100%	0%
Aug D	Time Stretch Slow, Time Stretch Fast	10.71%	231	100%	0%
Aug E	Volume Up, Volume Down	14.29%	288	100%	0%
Aug F	Small Noise, Medium Noise, Background Noise, Pitch Up, Pitch Down	85.71%	277	30.43%	39.29%
Aug G	Small Noise, Medium Noise, Background Noise, Pitch Up, Pitch Down, Time Stretch Slow, Time Stretch Fast	92.86%	285	43.48%	14.29%
Aug H	Small Noise, Medium Noise, Background Noise, Pitch Up, Pitch Down, Time Stretch Slow, Time Stretch Fast, Volume Up, Volume Down	100%	333	34.78%	32.14%

In Aug B, noise augmentation improved robustness against environmental disturbances; however, the FAR and FRR values remained relatively high compared to Aug A. Meanwhile, Aug C, Aug D, and Aug E produced extremely poor performance, with validation accuracy close to 10% and FAR reaching 100%. These results indicate that dominant single type augmentation caused excessive distortion of speaker characteristics, making it difficult for the model to learn the original voice patterns effectively.

Aug F and Aug G achieved relatively high accuracy; however, the increasing number of augmentation variations significantly increased the FAR values. This finding indicates that excessive augmentation may increase embedding overlap among speakers, causing the system to become more permissive toward unknown speakers.

In Aug H, the model achieved a test accuracy of 100%; however, the FAR remained relatively high at 34.78%. In biometric authentication systems, FAR is considered more

critical than accuracy because it directly reflects the risk of unauthorized user acceptance. Therefore, high accuracy does not necessarily indicate a secure authentication system.

Based on the overall experimental results, Aug A was selected as the optimal augmentation configuration because it provided the best balance between accuracy and security performance. The results demonstrate that lightweight augmentation is more effective for limited datasets compared to excessive augmentation strategies, which may lead to over augmentation and reduced speaker embedding stability.

E. Best Model Selection

Based on all stages of hyperparameter tuning and audio augmentation analysis, this study obtained the best configuration for the final open set speaker verification model. The optimal configuration consisted of MFCC-C (N_MFCC = 40, N_FFT = 1024, HOP_LENGTH = 256, N_MELS = 40), CNN-E with three convolution blocks (32–64–128), an embedding dimension of 64, and lightweight augmentation using noise, pitch shifting, and time stretching with one variation for each augmentation type.

The configuration was selected based on the balance between classification performance and biometric security performance. The experimental results showed that the selected model achieved a test accuracy of 96.43%, a FAR of 8.7%, and an FRR of 28.57%. The low FAR value indicates that the system was capable of effectively rejecting unknown speakers, thereby minimizing the risk of unauthorized access. In biometric authentication research, FAR is considered one of the most important evaluation parameters because it is directly related to system security. However, the obtained FAR value also indicates that several unauthorized speakers were incorrectly accepted by the system. In practical biometric security applications, such false acceptance may increase the risk of unauthorized access. Nevertheless, the current study prioritizes lightweight implementation and limited scale evaluation rather than production level security deployment, while future work will focus on reducing FAR through larger scale datasets, improved speaker embeddings, and antispoofing evaluation under more diverse real world conditions.

Furthermore, the analysis demonstrated that excessive augmentation did not necessarily improve system performance. In several scenarios, excessive augmentation increased embedding overlap among speakers, resulting in significantly higher FAR values. This condition indicates that overly aggressive augmentation may distort the speaker feature distribution, causing the system to become more permissive toward unknown speakers. Therefore, lightweight and controlled augmentation was considered more suitable for limited datasets, such as those used in this study.

Figure 8 illustrates the training accuracy and validation accuracy curves of the best model configuration. The curves demonstrate that the training process was relatively stable without severe overfitting. The increasing training accuracy was followed by stable validation accuracy until the training process stopped at epoch 45 using the early stopping mechanism.

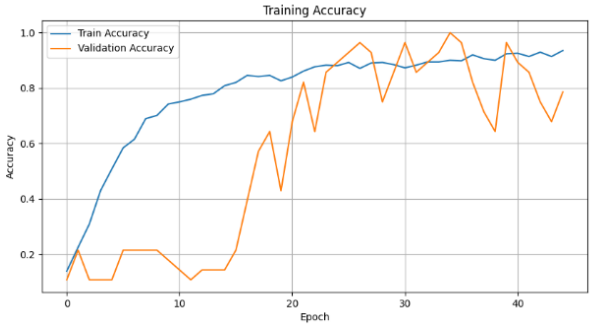

Figure 8. Training Accuracy Best Model Selection

Figure 9 presents the training loss and validation loss curves. The results show that the loss gradually decreased during the training process, indicating that the model successfully learned representative speaker features. The stable validation loss also demonstrates that the model achieved relatively good generalization capability on unseen data.

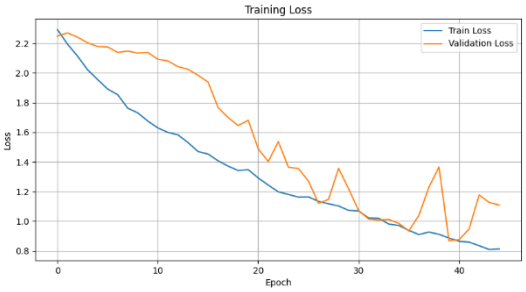

Figure 9. Training Loss Best Model Selection

The system security evaluation is presented in Figure 10 using the FAR and FRR curves against varying threshold values. The curves illustrate the trade off between false acceptance and false rejection rates. Based on the experimental results, the optimal threshold was obtained at 0.95, which provided the best balance between FAR and FRR. Threshold selection is an essential stage in open set speaker verification because it directly affects the sensitivity of the system in accepting or rejecting users.

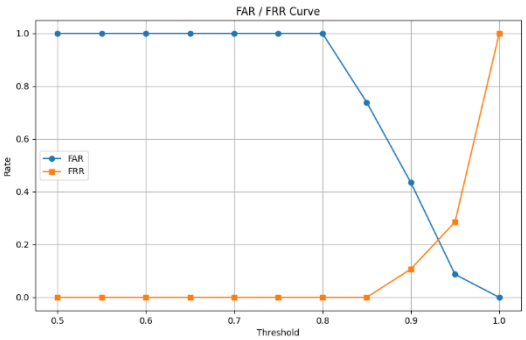

Figure 10. FAR and FRR Comparison Best Model Selection

Figure 11 presents the confusion analysis showed that the proposed speaker verification system achieved 20 true positive cases and 21 true negative cases, indicating that most registered speakers were correctly accepted while most unknown speakers were successfully rejected. Only 2 false positive cases were observed, demonstrating good security performance against unauthorized access. However, the system still produced 8 false rejection cases, indicating that several genuine speakers were incorrectly rejected due to the relatively strict decision threshold. The obtained results achieved a security accuracy of 80.39%, precision of 90.91%, recall of 71.43%, and F1-score of 80.00%, showing that the proposed CNN Embedding model provided promising performance for lightweight open set speaker verification systems.

	Predicted Positive	Predicted Negative
Actual Positive	TP: 20	FN: 8
Actual Negative	FP: 2	TN: 21

```
========== SECURITY METRICS ==========
Accuracy  : 0.8039 | Precision : 0.9091
Recall    : 0.7143  | F1-Score  : 0.8000
```

Figure 11. Security Confusion Matrix

Figure 12 presents t-SNE visualization was implemented to validate the separation of voice feature distributions between classes visually. The use of this algorithm aims to demonstrate that the embedding space generated by the proposed model has been well optimized, where voice samples from the same individual tend to form consistent clusters, while samples from different individuals or unregistered subjects (unknown speakers) are significantly separated by clear decision boundaries.

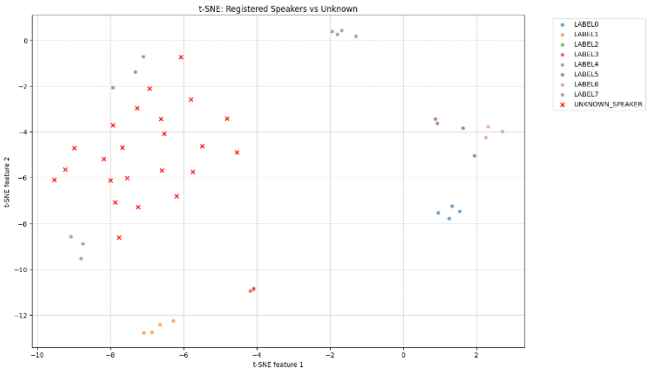

Figure 12. t-SNE Visualization of Best Model Selection

Figure 13 shows the confusion matrix of the speaker classification results on the testing dataset. Based on the confusion matrix, most voice samples were correctly classified into their corresponding speaker classes. This result demonstrates that the proposed CNN model effectively learned the voice characteristics of registered users.

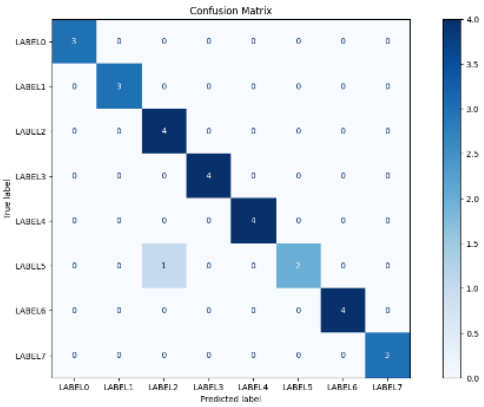

Figure 13. Confusion Matrix Best Model Selection

Overall, the experimental results demonstrate that the combination of hyperparameter optimization and lightweight augmentation significantly improved the performance of the open set speaker verification system. The best model configuration obtained in this study is considered sufficiently optimal for implementation in real time voice biometric security systems.

IV. CONCLUSION

This study investigated hyperparameter optimization for a CNN based open set speaker verification system using MFCC features and speaker embedding. The optimization process was conducted through several experimental stages, including MFCC parameter tuning, CNN architecture tuning, embedding dimension tuning, and audio augmentation analysis for voice biometric authentication.

Based on the MFCC tuning results, the MFCC-C configuration with parameters N_MFCC = 40, N_FFT = 1024, HOP_LENGTH = 256, and N_MELS = 40 showed stable biometric security performance with a FAR of 0% and an FRR of 7.14%. In the CNN architecture tuning stage, the CNN-E configuration with three convolution blocks (32–64–128) and a dropout rate of 0.35 provided a promising balance between classification accuracy, model complexity, and security performance. Furthermore, the embedding dimension analysis indicated that an embedding dimension of 64 produced favorable performance in maintaining speaker feature representation and computational efficiency.

The audio augmentation analysis showed that lightweight augmentation using a combination of noise injection, pitch shifting, and time stretching with limited variations provided more stable performance compared to more complex augmentation strategies. Excessive augmentation on limited datasets tended to increase embedding overlap among speakers, which may contribute to higher False Acceptance Rate (FAR) values.

Overall, the experimental results indicate that hyperparameter optimization influenced the performance of the proposed open set speaker verification system, particularly in balancing classification accuracy and

biometric security metrics such as FAR and FRR. The obtained CNN-Embedding configuration demonstrated promising performance for lightweight and limited scale speaker verification applications. However, the dataset used in this study contained limited speaker diversity and controlled recording conditions, and no statistical significance testing was conducted across experimental configurations. Therefore, further evaluation using larger scale datasets, more diverse recording environments, and statistical validation methods is required to better assess the generalization capability of the proposed system.

REFERENCES

[1] I. P. Ihsan, S. Buwarda, H. Novianty, I. A. Putra, and U. Fajar, "Voice Recognition Untuk Otomatisasi Sistem Pengakses Pintu," *JSAI : Journal Scientific and Applied Informatics*, vol. 4, no. 01, 2021, doi: 10.36085.

[2] H. Isyanto, A. S. Arifin, and M. Suryanegara, "Fast and Accurate Voice Biometrics with Deep Learning Algorithm of CNN Depthwise Separable Convolution Model and Fusion of DWT-MFCC Methods," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 8, no. 3, p. 431, Oct. 2022, doi: 10.26555/jiteki.v8i3.24515.

[3] Haris Isyanto, "Accurate, Fast and Low Computation Cost of Voice Biometrics Performance using Model of CNN Depthwise Separable Convolution and Method of Hybrid DWT-MFCC for Security System," *Buletin Pos dan Telekomunikasi*, vol. 22, no. 1, Jun. 2024, doi: 10.17933/bpostel.v22i1.393.

[4] A. Koriah, P. Teknik Informatika, S. N. Syaikh Zainuddin Anjani Jalan Raya Mataram, and L. Timur, "Rancang Bangun Sistem Keamanan Pintu Rumah Dengan Voice Recognition Dan Rfid Gelang Berbasis Iot (Design A Home Door Security System With Voice Recognition And Iot-Based Rfid Bracelets)." doi: https://doi.org/10.46764/teknimedia.v5i2.241.

[5] F. A. Alaba, M. Othman, I. A. T. Hashem, and F. Alotaibi, "Internet of Things security: A survey," *Journal of Network and Computer Applications*, vol. 88, pp. 10–28, 2017, doi: https://doi.org/10.1016/j.jnca.2017.04.002.

[6] V. Muthumanikandan, Shajeth, and V. Sathya, "Voice-driven IoT: Revolutionizing home and hospital automation for enhanced security," *AIP Conf. Proc.*, vol. 3383, no. 1, p. 040010, Feb. 2026, doi: 10.1063/5.0308829.

[7] A. B. Arief, "Perancangan Smart Home Berbasis Internet Of Things Dengan Fokus Pada Pengendalian Suara Melalui Integrasi Google Home," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 14, no. 1, Jan. 2026, doi: 10.23960/jitet.v14i1.8872.

[8] Mutiara Syafrizal, Vidya Ikawati, Agus Siswanto, and N. Lestari, "Sistem Keamanan Door Lock Berbasis Voice Recognition Dengan Natural Language Processing," *Infotronik : Jurnal Teknologi Informasi dan Elektronika*, vol. 9, no. 1, pp. 1–11, Jun. 2024, doi: 10.32897/infotronik.2024.9.1.3611.

[9] X. Liu, M. Sahidullah, and T. Kinnunen, "Learnable MFCCs for Speaker Verification," Feb. 2021, [Online]. Available: http://arxiv.org/abs/2102.10322

[10] M. Ardiana, T. Dutono, and T. B. Santoso, "Gender Classification Based Speaker's Voice using YIN Algorithm and MFCC," in *2021 International Electronics Symposium (IES)*, 2021, pp. 438–444. doi: 10.1109/IES53407.2021.9593959.

[11] M. Ardiana, T. Dutono, D. Tri, and B. Santoso, "Jurnal Politeknik Caltex Riau Identifikasi Jenis Kelamin Secara Real Time Berdasarkan Suara Pada Raspberry Pi," 2022. [Online]. Available: https://jurnal.pcr.ac.id/index.php/jkt/

[12] A. Ashar, M. Shahid Bhatti, and U. Mushtaq, "Speaker Identification Using a Hybrid CNN-MFCC Approach," 2020. doi: 10.1109/ICETST49965.2020.9080730.

[13] A. Wirdiani, S. Ndung'u Machetho, K. Gede, D. Putra, R. S. Hartati, and H. A. Ferdian, "Improvement Model for Speaker Recognition using MFCC-CNN and Online Triplet Mining," vol. 14, no. 2, 2024.

[14] M. Kumar, N. Mohd, G. Shivam, A. Goyal, D. Parashar, and R. Khan, "Hybrid Aquila optimizer-Harris Hawks optimization for CNN hyperparameter tuning in brain tumor classification," *Sci. Rep.*, vol. 16, no. 1, Dec. 2026, doi: 10.1038/s41598-026-43329-7.

[15] C. Author, B. Kanata, and S. M. Al Sasongko, "Enhancing Heart Sounds Classification Using MFCC And CNN," *International Journal of Informatics and Computation (IJICOM)*, vol. 8, no. 1, 2026, doi: 10.35842/ijicom.

[16] S. Simboni Tege, K. Katalay Pierre, O. Oshasha Fiston, S. Frey, A. Ntumba Nkongolo, and B. Kuya Jirince, "Comparative Evaluation of MFCC and Mel-spectrogram Features for CNN-Based Respiratory Abnormality Detection," 2026. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[17] C. Author, B. Kanata, and S. M. Al Sasongko, "Enhancing Heart Sounds Classification Using MFCC And CNN," *International Journal of Informatics and Computation (IJICOM)*, vol. 8, no. 1, 2026, doi: 10.35842/ijicom.

[18] A. Gusev *et al.*, "Deep Speaker Embeddings for Far-Field Speaker Recognition on Short Utterances," Feb. 2020, [Online]. Available: http://arxiv.org/abs/2002.06033

[19] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333. doi: 10.1109/ICASSP.2018.8461375.

[20] G. Costantini, V. Cesarini, and E. Brenna, "High-Level CNN and Machine Learning Methods for Speaker Recognition," *Sensors*, vol. 23, no. 7, Apr. 2023, doi: 10.3390/s23073461.

[21] F. N. Rahman, T. Listyorini, and E. Supriyati, "Analisis Akurasi Cnn Pada Data Olah Suara Manusia Menggunakan Parameter Koefisien Mfcc Dan Max Length," 2025.

[22] H. Isyanto, A. S. Arifin, and M. Suryanegara, "Fast and Accurate Voice Biometrics with Deep Learning Algorithm of CNN Depthwise Separable Convolution Model and Fusion of DWT-MFCC Methods," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 8, no. 3, p. 431, Oct. 2022, doi: 10.26555/jiteki.v8i3.24515.

[23] C. Li *et al.*, "Deep Speaker: an End-to-End Neural Speaker Embedding System," May 2017, [Online]. Available: http://arxiv.org/abs/1705.02304

[24] J. Galić, B. Marković, Đ. Grozdić, B. Popović, and S. Šajić, "Whispered Speech Recognition Based on Audio Data Augmentation and Inverse Filtering," *Applied Sciences (Switzerland)*, vol. 14, no. 18, Sep. 2024, doi: 10.3390/app14188223.

[25] S. Seo and J. H. Kim, "Self-attentive multi-layer aggregation with feature recalibration and deep length normalization for text-independent speaker verification system," *Electronics (Switzerland)*, vol. 9, no. 10, pp. 1–14, Oct. 2020, doi: 10.3390/electronics9101706.