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

Haris Isyanto[1]*, Wahyu Ibrahim[2], Riza Samsinar[3]
[1,2,3] Department of Electrical Engineering, Faculty of Engineering, Universitas Muhammadiyah Jakarta, Indonesia
Cempaka Putih Tengah 27 Jakarta Pusat 10510
*Email : [1]haris.isyanto@umj.ac.id

ARTICLE INFORMATION

Received on 1 April 2024
Revised on 11 June 2024
Accepted on 25 June 2024

Keywords:
CNN Depthwise Separable Convolution
CNN Residual
Discrete Wavelet Transform
Mel-Frequency Cepstral Coefficients
Voice biometrics

A B S T R A C T

Identity theft, a pervasive criminal risk in the digital realm, particularly in online transactions, demands innovative security solutions. Voice biometrics, a cutting-edge technology, have been developed to ensure the protection of one's identification. This study, a significant step forward, focuses on the development of voice biometrics using deep learning, specifically CNN Depthwise Separable Convolution (DSC) and CNN Residual. The research on these two systems was conducted to determine accuracy, performance evaluation, computing load, and training process time for effectively, rapidly, and accurately verifying user voice for banking transaction security. The initial CNN residual test yielded a high validation accuracy of 98.6345%. However, the large number of CNN residual parameters resulted in a training time of 7.37 seconds, increasing the computational workload. The second CNN DSC test exhibited a high validation accuracy of 98.3542%. The CNN DSC was successful in decreasing the parameter count, resulting in a reduction of 5.12 seconds in training time. Upon analyzing the test results, it is clear that the CNN DSC has superior performance, resulting in faster training times and less memory consumption. This effectively addresses the problem of high computational costs and significantly enhances user identity security in banking transactions, a crucial aspect of modern banking.

1. Introduction

The escalating issue of fraudulent actions and the illegal acquisition of personal data, commonly referred to as identity theft, has become a major threat in the field of cybercrime. This problem is closely associated with the increasing frequency of Internet usage in various sectors, especially in online transactions made via the Internet network. Ensuring security is of utmost importance when transmitting digital data, especially in order to protect the confidentiality of one's identity (Kanervisto et al., 2022; Sarkar & Singh, 2020; Yusuf et al., 2020).

The current identifying methods still rely on traditional forms of identification, such as passwords, magnetic ID cards, and personal identifying number (PIN) codes on credit and debit cards. Yet, these approaches are not without their flaws, including the possibility of forgetfulness, physical damage to the card, loss, theft, and unauthorized access to the card's information. It's clear that we need more secure alternatives to protect our identities and information. (Arora & Bhatia, 2022; Filho et al., 2022; G. Singh et al., 2021; Wells & Usman, 2023).

In order to address this issue of criminal activity, a biometric technique was devised to authenticate individuals based on their unique biological traits obtained from their own bodies (Tyagi et al., 2019). Biometric-

based personal identification techniques have been specifically created as an alternative for high-level security access applications, such as government or military facilities, safeguarding access to sensitive data or information, financial transactions, and theft prevention (Rui & Yan, 2019).

A voice biometrics identification security approach was devised to tackle this difficulty, offering personalized safeguarding and guaranteeing protection against identity theft. Voice biometrics is a method that utilizes the specific patterns of the human voice to verify and validate individuals by analyzing their unique voice characteristics (Sholokhov et al., 2020; Yusuf et al., 2020).

Integrating voice biometrics does not need specialized hardware and is a more economical option compared to other biometric methods, such as fingerprint readers or retinal scanners. Voice biometrics offers enhanced security, user-friendly features, and very precise recognition capabilities for individual authentication. Speech biometrics utilizes user speech commands to transmit voice messages over internet-connected computers and cellphones, eliminating the necessity of typing a password. The voice instructions inputted into the voice biometrics system are compared to the stored voices in the database (Sholokhov et al., 2020).

Prior studies in speech biometrics mainly employed machine learning techniques to analyze data for voice object recognition. However, more studies are needed to explore the utilization of deep learning methods, namely convolutional neural networks (CNN), in voice biometrics. The primary challenges associated with machine learning are the limitations in handling vast quantities of data and the intricacy of data processing when it comes to the recognition or identification of objects in pictures and voices (Isyanto et al., 2022)

CNN is a sophisticated deep learning system specifically designed to effectively analyze large volumes of complex input for the purpose of object recognition (Li et al., 2022). Adding advanced techniques like the CNN Residual and CNN Depthwise Separable Convolution (DSC) models to the CNN Standard model is needed to improve the accuracy and performance of the voice biometrics system. The advantage of CNN Residual and CNN DSC compared to standard CNN is that the level of accuracy of the performance of both CNNs in predicting voice biometric classification would improve access to the user's identity security system. The CNN residual is employed to optimize the training and validation process while also improving the accuracy of classification (Ihsanto et al., 2020b; Liu, 2022; 小川, 2021). Conversely, the CNN DSC is utilized to reduce the number of parameters and mathematical operations required in convolution procedures. The benefits of reducing parameters and computational complexity using CNN DSC result in faster training times and less memory usage, allowing you to solve problems with high computational costs (Ihsanto et al., 2020a; Jung et al., 2021; Lu et al., 2022; Shan et al., 2021).

The researcher's contribution is to develop a voice biometrics scheme framework that uses the CNN DSC model's deep learning algorithm to conduct a comparative performance analysis with the CNN residual model. This study develops a voice biometrics system that efficiently, quickly, and accurately uses algorithms to identify, verify, and analyze the voice patterns of different users. The aim of developing a voice biometric framework scheme in the user verification process using the CNN DSC and CNN residual models is to determine the level of accuracy, evaluate the performance of the computational load, and determine the time the training process takes to predict voice classification in banking security system access. The level of accuracy of CNN's performance in predicting voice biometric classification would improve the user's identity security system. This research is expected to significantly increase the accuracy of prediction classification performance to more than 90%.

2. Literature review

2.1. Literature review

Prior studies have only partially explored the subject of voice biometrics, specifically concentrating on the amalgamation of speech recognition and speaker recognition. Academic publications generally concentrate on voice recognition and speaker recognition as separate subjects. The research pertinent to this topic is presented in Table 1.

A precision of around 76% was attained in the fields of speech recognition and speaker recognition by employing machine learning methods and feature extraction.

The fields of speech recognition and speaker recognition utilize deep learning and feature extraction techniques. The degree of accuracy attained using these methods ranges from 71% to 90%.m 71% to 90%.

Different machine learning methods, like k-Nearest Neighbors (k-NN) (M S & P S, 2021), Support Vector Machines (SVM) (Batista et al., 2020; Chowdhury & Ross, 2020; Nainan & Kulkarni, 2021; Sen et al., 2021), and feature extraction using Mel-frequency cepstral coefficients (MFCC) (Chowdhury & Ross, 2020; Hidayat & Winursito, 2020), have been used in the past to study voice biometrics. In addition, researchers have also used Gaussian Mixture Models (GMM) in combination with MFCC feature extraction (Sen et al., 2021). No study on voice biometrics has been discovered that specifically investigates the application of deep learning techniques.

Table 1. Literature review

Types of Voice Patterns	Reference	Algorithm Models	Extraction Features	Explanation
Speech Recognition	(Batista et al., 2020)	Machine Learning (ML) SVM	MFCC	Machine Learning (ML) has limited capacity to handle large datasets and is less adept at processing intricate data. Additionally, Mel-Frequency Cepstral Coefficients (MFCC) are susceptible to noise interference.
	(Huang et al., 2020; Ping, 2021)	Machine Learning HMM	MFCC	Same explanation as above.
	(Nayana et al., 2017; Sen et al., 2021)	Machine Learning GMM	i-vector, MFCC	Same explanation as above.
	(Jolad & Khanai, 2022; M. K. Singh, 2023)	Deep Learning (DL) ANN	x	The computing capabilities of DL ANN are not very trustworthy. Lacks an extraction capability.
	(Alsobhani et al., 2021; Chai et al., 2021)	Deep Learning DNN	X	Same explanation as above.
	(Hao et al., 2021)	Deep Learning RNN dan LSTM	X	Deep Learning Recurrent Neural Networks (DL RNN) exhibit reduced compatibility compared to Convolutional Neural Networks (CNN). However, the Long Short-Term Memory (LSTM) technique can be employed to address the issue of gradient loss in RNNs. However, it lacks an Extraction Feature.
	(Alsobhani et al., 2021; Taye, 2023)	Deep Learning CNN	x	Large-scale data problems and complicated data processing are both handled well by DL CNN. However, it lacks an extraction feature.
Speaker Recognition	(Malik et al., 2020)	Machine Learning GMM	i-vector, MFCC, PNCC, RASTA PLP	The limitation of machine learning is its limited capacity to handle large volumes of data. Additionally, its extraction feature suffers from noise resistance.
	(M S & P S, 2021; Nainan & Kulkarni, 2021)	Machine Learning SVM-GMM	X	Same explanation as above. Does not have an Extraction Feature.

Types of Voice Patterns	Reference	Algorithm Models	Extraction Features	Explanation
Voice Biometrics	(Huang et al., 2020; Wei, 2020)	Machine Learning HMM	i-vector and MFCC	Same explanation as above.
	(M. K. Singh, 2023)	Deep Learning ANN	LPC, MFCC, ZCR	A lower level of computing reliability is exhibited by DL ANN. The issue with this extraction capability is that it isn't noise-resistant.
	(Quang et al., 2021)	Deep Learning DNN	i-vector and MFCC	Same explanation as above.
	(Andra & Usagawa, 2021; M. K. Singh, 2023)	Deep Learning RNN	X	DL RNN has a gradient loss issue, and it is less compatible than CNN. lacks the ability to extract.
	(Chowdhury & Ross, 2020; M. K. Singh, 2023)	Deep Learning CNN	MFCC	Large-scale data problems and complicated data processing are both handled well by DL CNN. The MFCC is noise-sensitive.
	(Moreno & Lopes, 2018)	Machine Learning k-NN	x	Only modest amounts of data can be processed by ML, while complicated data can be processed less well. lacks the ability to extract.
	(Duraibi et al., 2020; Singla & Verma, 2023)	Machine learning SVM	MFCC	Only modest amounts of data can be processed by ML, while complicated data can be processed less well. The MFCC is noise-sensitive.
	(Amjad Hassan Khan & P. S. Aithal, 2022; Pawade et al., 2022)	Machine learning GMM	MFCC	Same explanation as above.
	The 1st scheme developed in this research	Deep-learning CNN Residual Model	Hybrid DWT-MFCC	It can greatly improve classification accuracy by streamlining the training and validation procedures with CNN residual.
	The second scheme was developed in this research	Deep-learning CNN Depthwise Separable Convolution (DSC) Model	Hybrid DWT-MFCC	The use of CNN DSC enables a reduction in the number of parameters and arithmetic operations involved in convolution operations, hence decreasing the computing load during training. It may effectively address the issue of expensive computational requirements and enhance the speed of prediction processing time in classification.

The goal of this study is to improve the efficiency of voice biometrics by utilizing deep-learning CNN algorithms, specifically CNN DSC and CNN residual, within a given schematic framework. CNN's performance is capable of effectively addressing challenges related to handling extensive data and complex data structures in the process of user voice identification. In order to enhance the performance of the CNN model, it is optimized using CNN residual and CNN DSC techniques. This CNN residual model can streamline training and validation procedures while simultaneously improving classification accuracy. The CNN DSC model solves the problem of computing costs that are too high and speeds up prediction processing in speech biometrics classification beyond what a regular CNN can do. The benefit of CNN DSC lies in its ability to decrease the number of training parameters, minimize arithmetic operations in convolution processes, and reduce the computing burden during training. In addition, the DWT-MFCC hybrid extraction feature is employed, which effectively eliminates noise (denoising) in signal processing to enhance voice quality, identify voice patterns, extract features from voice characteristics, and isolate relevant voices from irrelevant ones.

This research proposal aims to develop a sophisticated framework for creating a speech biometric system. This framework will facilitate the design of accurate, fast, and practical algorithms for user voice identification

and verification/authentication. The primary purpose of these algorithms is to facilitate access to banking security systems. The goal of this study is to improve prediction categorization accuracy by a margin of over 90%.

2.2. Deep learning of CNN DSC model

The study presents the CNN Depthwise Separable Convolution (DSC) as an incremental enhancement to conventional convolutions (Conv layers). An advancement that has been made is the implementation of Depthwise Separable Convolution (DSC) (Shan et al., 2021). Previous research has shown that the DSC model is good at lowering the number of training parameters, reducing the amount of work that needs to be done on the computer, speeding up the process, and lowering the higher costs of computing that come with convolution operations compared to standard convolution (Shan et al., 2020). Figure 1 illustrates a comparison between standard convolution and DSC (Shan et al., 2021).

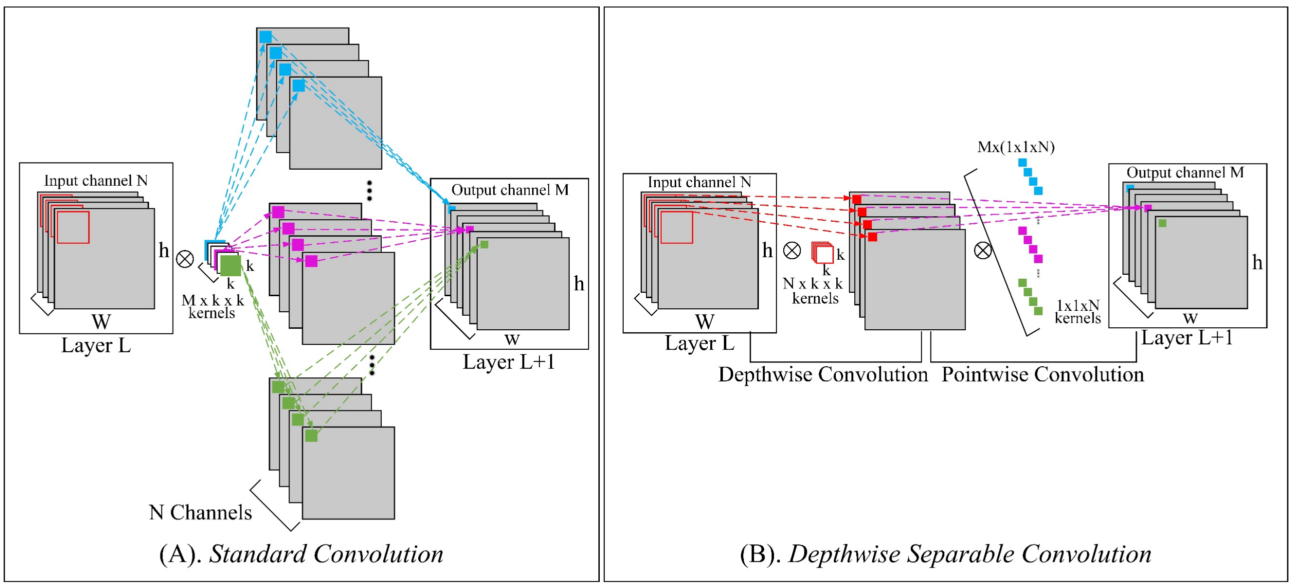


Figure 1. Depthwise separable convolution model

To be more specific, DSC divides ordinary convolution computations into depthwise and pointwise convolutions. Equation (1) in the text depicts the standard convolution computations for a solitary layer. Additionally, equations (2) and (4) serve as examples of the categorization of these computations into two distinct categories. (1) has three accumulation layers, specifically k, l, and m, that are used for doing traditional convolutional calculations. Equation (2) includes only one accumulation layer represented as "m," but equation (3) encompasses two accumulation layers, specifically "k" and "l." The equation (4) (Pyykkönen et al., 2020; Shan et al., 2021) illustrates the amalgamation of pointwise convolution and depthwise convolution in DSC calculations. Although DSC can efficiently reduce the number of arithmetic operations and computational burden, the results of feature map calculations are almost indistinguishable from those obtained using conventional convolution methods (Lu et al., 2022; Pyykkönen et al., 2020).

$$Conv(\omega, y)_{(i,j)} = \sum_{k,l,m}^{K,L,M} \omega_{(k,l,m)} \cdot y_{(i+k,j+l,m)} \quad \ldots 1)$$

$$PointwiseConv(\omega, y)_{(i,j)} = \sum_{m}^{M} \omega_m \cdot y_{(i,j,m)} \quad \ldots 2)$$

$$DepthwiseConv(\omega, y)_{(i,j)} = \sum_{k,l}^{K,L} \omega_{(k,l)} \odot y_{(i+k,j+l)} \quad \ldots 3)$$

$$\text{SepConv}(\omega_p, \omega_d, y)_{(i,j)} = \text{PointwiseCon}v_{(i,j)}\left(\omega_p, \text{DepthwiseConv}_{(i,j)}(\omega_d, y)\right) \quad \text{.....4)}$$

2.3. Deep learning of CNN Residual model

A residual shortcut is a technique employed in convolutional neural networks (CNNs) to establish connections across different levels within a branch, allowing for the creation of further branch layers. By utilizing the residual technique, it is feasible to decrease the length of training repeats while concurrently improving the accuracy value. To achieve this improvement, one can increase the number of parameters, filters, and layers. The following equation is the comprehensive formula for the shortened residual identity function, as seen in Equation (5) (Ihsanto et al., 2020b).

$$y = F[x, (W_i)] + x \quad \text{.....5)}$$

The feature map obtained after applying residuals is labeled as y. The filter, which represents the residual mapping, is denoted as F [x, (Wi)]. Here, x corresponds to the input feature map. The layer group named "Wi" is exempted from dimension modifications for the x and y axes when performing operations such as down sampling or up sampling. The provided input consists of the list (Ihsanto et al., 2020b; 小川, 2021).

3. Method

3.1. Preprocessing voice datasets

The method of generating speech data begins with the user's input voice, which is captured by the microphone of a laptop device utilizing headphones for recording. This voice dataset was generated using a total of 10 people, identified as Voice User0 to Voice User9 (VU0–VU9). Each VU input consists of a speaker's voice and speech (spoken words). Every participant completed a voice sample by speaking Indonesian. The act of producing voice was performed in an enclosed space specifically designed for conducting experiments with electricity. Table 2 displays the preprocessing table for voice datasets.

Table 2. Preprocessing voice datasets

User	Voice Input	Voice Content	Language	Environment
VU0	Microphone device laptop using headphones	Voice data input: Speaker and Speech	Speaking in Indonesian	Indoor (Electrical Laboratory)
VU1				
VU2				
VU3				
VU4				
VU5				
VU6				
VU7				
VU8				
VU9				

The production of voice data starts with the voice user's input through the microphone device. Every VU input contains speakers and speech. Each VU fills a sample of voices by speaking in Indonesian. Then, change the audio sample file format to the WAV file type format. This WAV format is the best file format for the voice sample, and then all of the user's sample data sets are converted one by one alternately on 10 users. To create a uniform voice sample file in the production of voice data, you need to set the following parameters: First, set the stereo voice type to mono voice so that the voice produced on the mono audio is heard more clearly on the recording of the voice samples. Second, a sample rate of voice frequency should be set to 16,000 Hz. Third,

truncate silence is used to eliminate the user's voice pause based on the above sample. The process of removing the voice pauses above then obtains a no-pause sample of the voice. The sample is then segmented to form a segmentation file, which cuts the duration of each user's sample to seconds, thus forming a sample file. So, the results obtained are collected. The voice file will be saved into a segmented voice data folder based on the result of the segmentation process. The voice file, as a voice data set, is prepared to be trained on the deep learning algorithm of the CNN model, both on the CNN Residual and CNN DSC models. The following flowchart for preprocessing voice datasets can be seen in Figure 2.

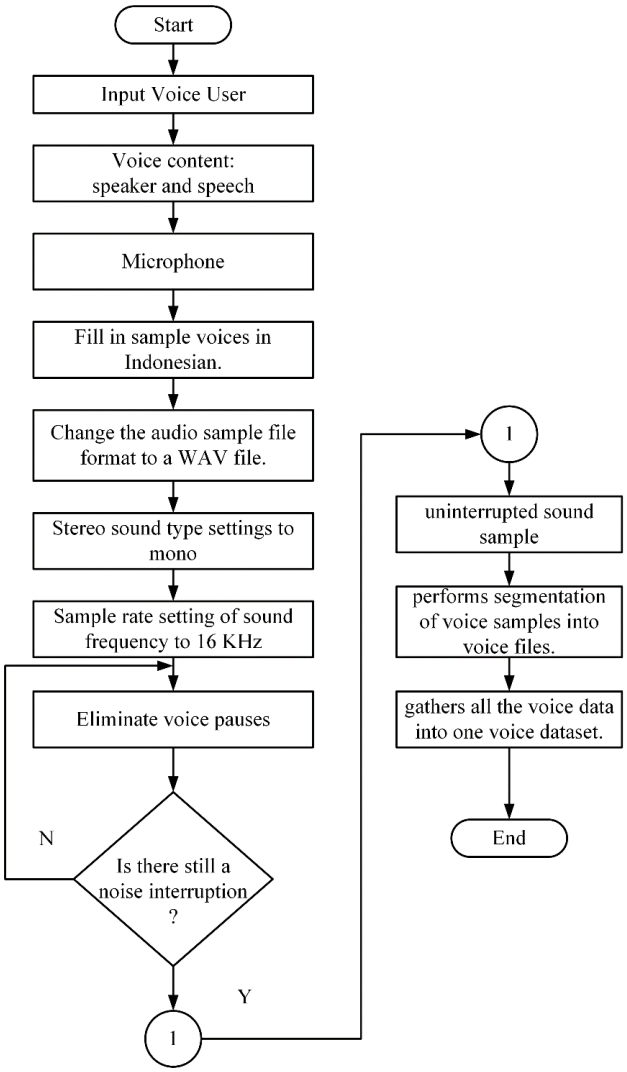


Figure 2. Flowchart of preprocessing voice dataset

Following the preprocessing of voice data, segmentation is performed by dividing audio recordings into one-second intervals for each user's voice. This procedure is carried out for a total of 10 users. The outcome corresponds to the count of segmentations in 3,000, 6,000, 9,000, 12,000, and 15,000 audio files, all of which have a sampling rate of 16,000 Hz, it could be seen in Table 3.

Table 3. User voice data set

User	Number of voice sample segmentation (file)					*Sample rate* (Hz)
VU0	300	600	900	1,200	1,500	16,000
VU1	300	600	900	1,200	1,500	
VU2	300	600	900	1,200	1,500	
VU3	300	600	900	1,200	1,500	
VU4	300	600	900	1,200	1,500	
VU5	300	600	900	1,200	1,500	
VU6	300	600	900	1,200	1,500	
VU7	300	600	900	1,200	1,500	
VU8	300	600	900	1,200	1,500	
VU9	300	600	900	1,200	1,500	
Total voice datasets (file)	**3,000**	**6,000**	**9,000**	**12,000**	**15,000**	

3.2. Development of CNN DSC and CNN Residual model with DWT-MFCC hybrid feature extraction

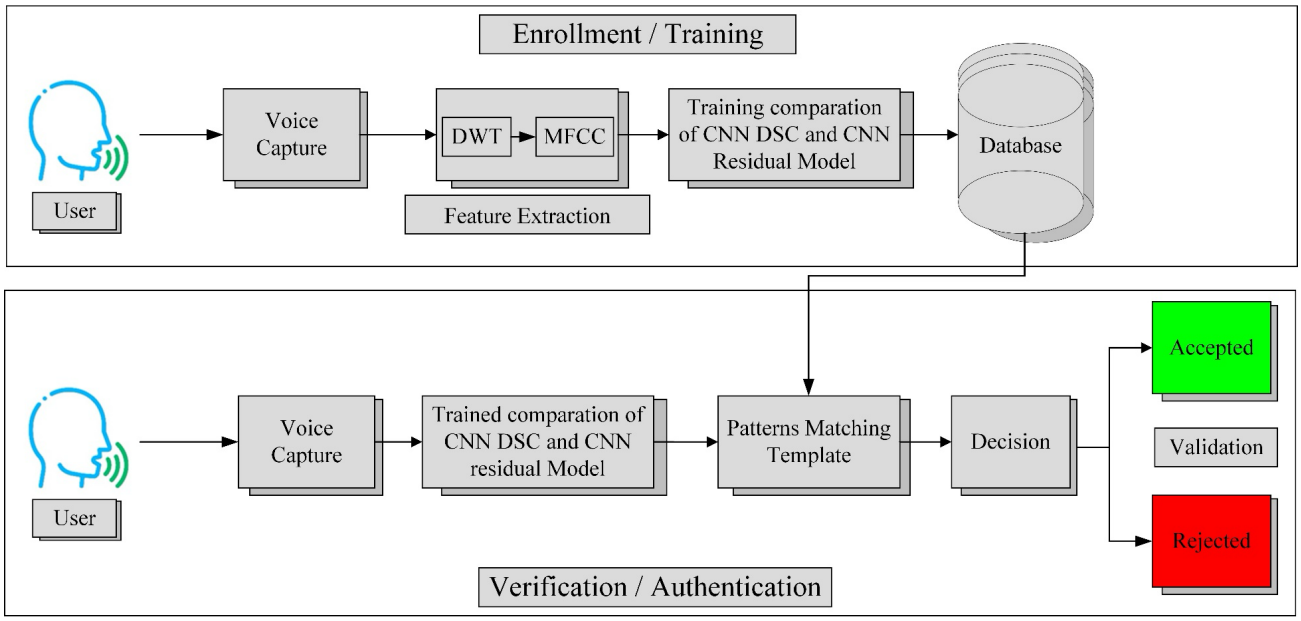


Figure 3. Development of a voice biometrics framework scheme for the processes of enrolment / training and verification/authentication user using the CNN DSC and CNN Residual models

The figure 3 illustrates the architectural design of the voice biometrics scheme that was created in this research. For user registration and training, the DWT-MFCC method is usually used to get the features, and the CNN residual and depthwise separable convolution (DSC) models are used for training. This training method involves acquiring capability by training the CNN model to recognize and classify user speech datasets. After the training phase is completed, the convolutional neural network (CNN) model will generate a trained CNN model. The CNN model that has been trained will thereafter be employed for the process of user verification. The user verification procedure entails classifying and authenticating speech datasets. This user verification method will be directly implemented on newly acquired speech data within the Trained Convolutional Neural Network (CNN) Model. The system will authenticate the voice data by comparing the user identification of the new direct voice data with the registered voice data in the database. Subsequently, the system will generate forecasts using the prediction precision of the data that has been trained on the Trained CNN Model. The categorization of the trained CNN model is optimized with the goal of maximizing prediction performance in

order to attain a high level of accuracy. The Trained CNN Model categorization produces verification and authentication results for the user's voice, indicating whether the voice data is legitimate or not, or if the user's voice data is accepted or denied. The obtained results consist of a comparison between CNN depthwise separable convolution and CNN residual models. This comparison includes data on training time, speaker recognition performance (accuracy and precision), speech recognition accuracy, and user response time during verification.

3.3. Architectural comparison of CNN residual and CNN depthwise separable convolution (DSC) algorithms

CNN Standard is a very efficient deep learning algorithm technology utilized for training and testing databases. The CNN Standard method is anticipated to enhance performance significantly compared to earlier machine learning algorithms. The CNN Residual and depthwise separable convolution architecture used in this research is illustrated in figure 4 and figure 5.

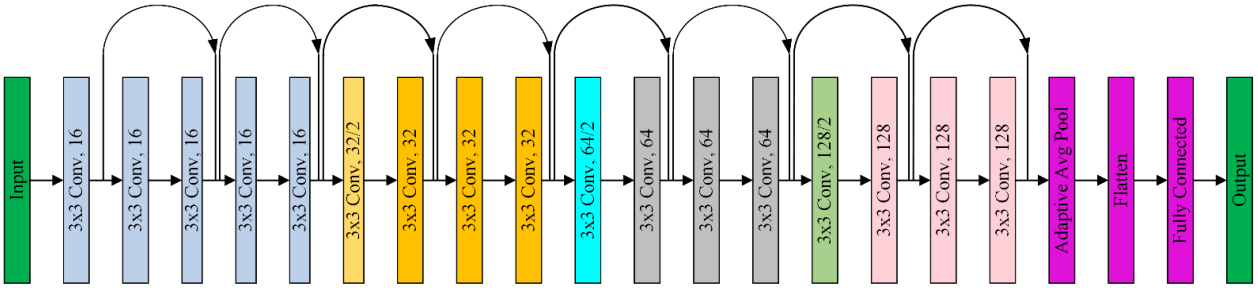


Figure 4. Architecture of CNN Residual

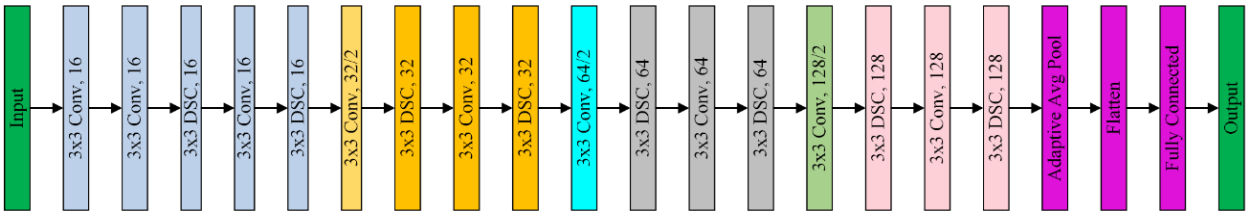
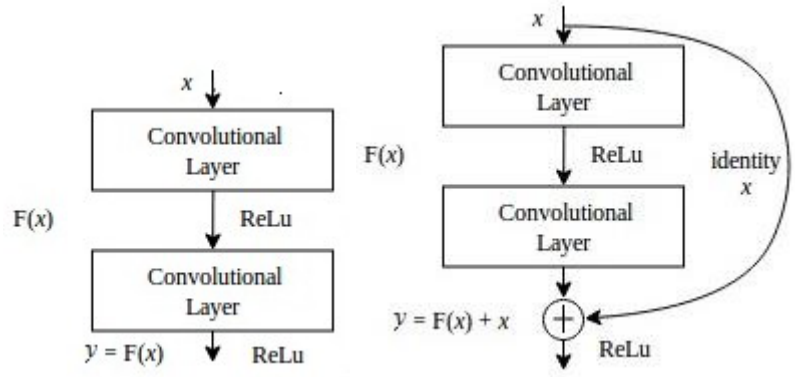


Figure 5. Architecture of CNN Depthwise Separable Convolution (DSC)

The CNN Residual architecture incorporates eight shortcut levels with a residual connection, with each layer being skipped over every two layers. This architecture consists of a total of 22 layers.

A CNN residual scheme design was created to enhance the performance of CNN Standard in voice biometrics systems. A CNN residual refers to a block or unit that contains skip connections, which are also known as identity connections.

The term "CNN Residual" refers to the base mapping represented by the variable y. The layers consist of mappings of F(x):= y − x that are placed on top of one another. The mapping is transformed into the sum of F(x) and x. Previous research indicates that optimizing residual mapping is more straightforward. The equation F(x)+x may be implemented using either feedforward neural networks or shortcut connections, as seen in Figure 6.



(a) (b)

Figure 6. (a) CNN Standard, (b) CNN Residual with identity connections (Ihsanto et al., 2020b; Isyanto et al., 2022)

Identity shortcut links do not include any extra layers or computational complexity. The complete network may still be educated in a continuous manner and can be readily adopted without requiring any alterations. The residual function, denoted as y−x (under the assumption that the input and output have the same dimensions), It is anticipated that the stacked layers will gradually converge towards the value of y, and subsequently, this layer will converge towards the residual function F(x) = y - x. The original function is modified by adding the value of x to F(x) (Ihsanto et al., 2020b; Isyanto et al., 2022)

The goal of this shortcut link is to mitigate the issue of vanishing gradients and minimize mistakes in the CNN residual network. This is a consequence of training mistakes, which are effectively minimized to far lower levels and may be used to validate data more broadly compared to CNN Standard. These results demonstrate that the issue of deterioration may be effectively addressed, resulting in a significant improvement in accuracy through the utilization of shortcut connections. This comparison validates the efficacy of residual learning in the CNN system. This demonstrates that utilizing identification shortcuts can enhance the training process by improving many aspects. Due to the smaller validation loss, CNN Residual's performance is superior to CNN Standard's. The utilization of this CNN residual model is necessary in order to streamline the training and validation process while simultaneously enhancing the accuracy of classification.

The proposed architecture for CNN DSC is characterized by the use of eight DSC layers instead of a convolutional layer. This architectural design comprises a grand total of 22 layers.

The suggested architecture of this DSC CNN involves substituting eight convolutional layers with eight depthwise separable convolution (DSC) layers. The standard structure of the convolutional layer is made up of a batch normalization layer, a rectified linear unit (ReLU) layer, and 3x3 convolutional layer operations. In the next step, the convolutional layer structure is changed to a Depthwise Separable Convolution (DSC) layer in the CNN DSC architecture demonstrated in Figure 7. Through the development of DSC, the convolution layer size is reduced to a 1×1 conv layer. In the DSC process, there are 3×3 DSC layers, batch normalization layers, ReLU layers, 1×1 conv layer, batch normalization layer, and ReLU.

In order to enhance the performance of convolutional neural networks (CNNs), a technique called CNN DSC optimization was devised. The development of CNN DSC optimization aimed to successfully decrease the number of training parameters, minimize arithmetic operations in convolution processes, and reduce computing complexity. The CNN DSC model effectively minimizes memory operations, reduces computing burdens, and accelerates the user voice classification process. As a result, it is highly beneficial for reducing the costs associated with intensive computational workloads.

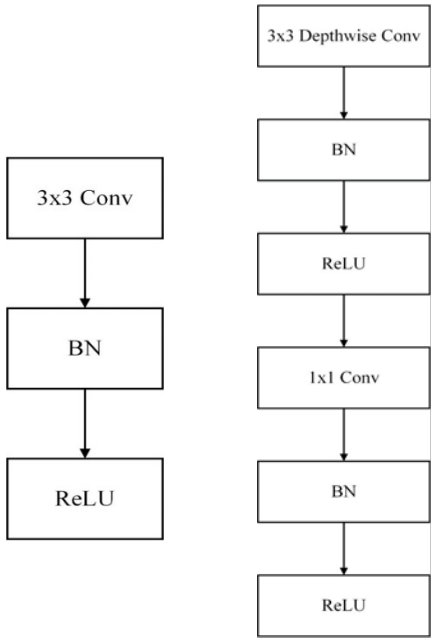


Figure 7. Differences in structure of building blocks (a) Standard Convolution dan (b) DSC

3.4. Evaluation of CNN model performance with Confusion matrix

A confusion matrix is a quantitative assessment tool used to evaluate the performance of machine learning and deep learning models in classification tasks. In order to assess the effectiveness of the classification system model, it is essential to compare the real values with the projected values. Utilizing a confusion matrix to assess system performance is critical in facilitating the algorithm training procedure for categorizing data labels or making precise predictions. This is intended to reduce the number of mistakes. When selecting the optimal system model from a range of deep learning models, it is crucial to take into account the confusion matrix technique.

When assessing the performance of a system model using a confusion matrix, it may be categorized into four distinct combinations of expected values and actual values. Table 2 displays four distinct outcomes of the categorization process: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The TN, FP, FN, and TP data collected will be utilized to calculate the accuracy and precision performance of the system model (Ibrahim et al., 2022; Isyanto et al., 2022) as a percentage. The following Evaluation of CNN model performance with Confusion matrix can be seen in table 4.

Table 4 is a description of the process of combining the confusion matrix.

- A true positive (TP) refers to positive data that is accurately identified as positive.
- A true negative (TN) refers to negative data that is accurately classified as negative.
- A false positive (FP) occurs when negative data is incorrectly anticipated as positive data.
- A false negative (FN) occurs when positive data is incorrectly forecast as negative.
- The predicted values are the program's output, which includes both positive and negative values.
- Actual values refer to values that are either true or false.

Table 4. Combination confusion matrix (Ibrahim et al., 2022)

		Actual value	
		Positive	Negative
Predicted value	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

The CNN model's performance is assessed by evaluating its outputs using a confusion matrix using speech biometric data. The accuracy and precision of data processing are determined by analyzing the actual values compared to the projected values using the number of voice sample files in the CNN DSC and CNN residual models. A performance evaluation was conducted on a set of voice file samples consisting of 3,000, 6,000, 9,000, 12,000, and 15,000 files. The accuracy and precision of the system model may be calculated using equations (1) and (2).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} x100\% \quad \text{6)}$$

$$Precision = \frac{TP}{TP + FP} x100\% \quad \text{7)}$$

4. Result and Discussion

4.1. Comparative Analysis of Voice Biometrics Training Performance Testing on Accuracy Validation and Loss Validation between CNN DSC and CNN Residual

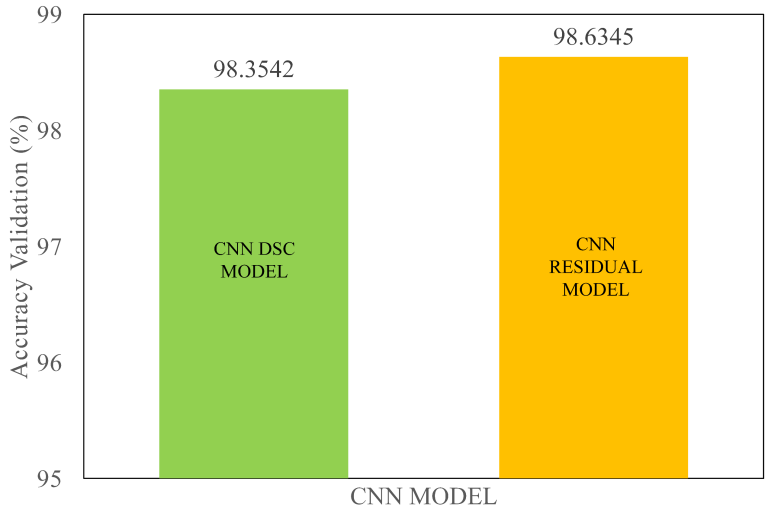


Figure 8. Performance comparing of Validation accuracy training on voice biometrics using CNN DSC and CNN residual model

Figure 8 illustrates a comparison of voice biometric training performance between the CNN residual and CNN DSC models. The results indicate that the CNN residual model fared better than the CNN DSC model in terms of validation performance, achieving a maximum accuracy of 98.6345%. A higher accuracy validation result indicates better performance on the CNN residual model. As the validation accuracy increases, the prediction performance of speech biometric classification also improves, enhancing the security system in user verification.

According to figure 9, the CNN residual model had a validation loss of 0.0501%, which was lower than the CNN DSC model's validation loss of 0.0759%. A smaller proportion of validation losses suggests superior performance in terms of loss validation on CNN Residual. As the validation loss performance decreases, the validation accuracy performance in speech biometric classification prediction increases, leading to enhanced security in user verification. Upon analyzing the study's findings comparing the CNN Residual and CNN DSC models, it can be concluded that CNN Residual's voice biometric training model exhibits superior validation accuracy compared to CNN DSC.

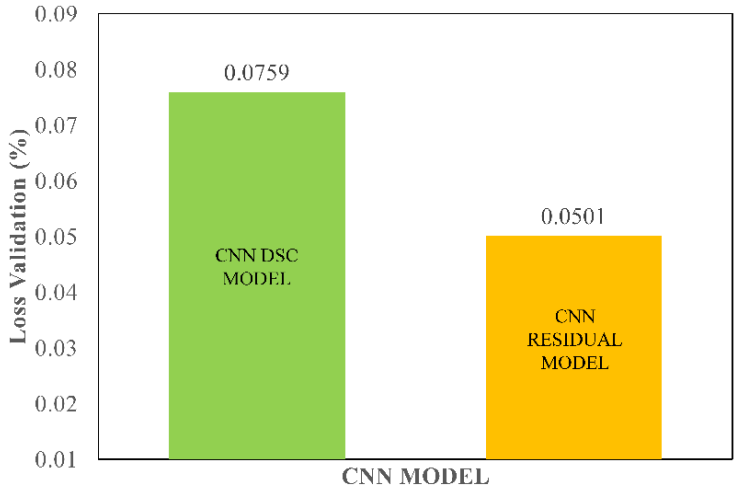


Figure 9. Performance comparing of Validation loss training on Voice biometrics using CNN DSC and CNN residual model

4.2. Comparative Analysis of Voice Biometrics Training Parameter Performance Testing between CNN Residual and CNN DSC Models

To assess the effectiveness of voice biometrics training parameters, the CNN residual and CNN DSC models must be compared using the DWT-MFCC approach. The goal of this test is to see how well the performance data works by looking at the training parameters of the CNN residual model and the CNN DWT model with a set of 15,000 voice samples. The test outcomes indicate that the CNN residual model includes 718,586 more parameters compared to the average CNN model, which has 364,506 parameters. The CNN residual model has a more significant number of parameters due to the inclusion of eight extra shortcut connections in its design. These connections allow for the skipping of every two layers in each connection. This shortcut link is included to mitigate the issue of vanishing gradients and training mistakes in the CNN residual. By using a shortcut link, the issue of deterioration may be effectively addressed, resulting in a significant improvement in the accuracy of speech categorization. The CNN residual parameter is 2.74 MB, which is larger than the CNN DSC model's size of 1.39 MB. Increasing the parameter size will have a direct impact on the quantity of data processed during the training stage of the CNN model. This will lead to higher memory needs and increased utilization of computational resources. Moreover, this will elongate the training procedure. According to the test results given earlier, it is clear that the CNN residual model's total parameter value is about twice as significant as that of the

CNN DSC model, with a difference of 354,080 parameters. The performance results of all parameter values will influence the CNN DSC model's training time, resulting in a faster training procedure compared to the CNN residual model.

Moreover, the parameter size value (parameter size) of the CNN residual model is about double the size of the CNN DSC model, with a difference of 1.35 MB. From the performance results of the parameter size, it can be deduced that decreasing it will result in a reduction in the computational burden during the training phase of the CNN DSC model. This will substantially reduce the costs involved with operating intricate computer systems. To examine the performance of training parameters in voice biometrics, one can refer to Table 5 for a comparison between the CNN residual and CNN DSC models.

Table 5. Performance Testing Comparison of Training Process Time on Voice Biometrics using the model of CNN DSC and CNN Residual

Parameter	Model CNN DSC	Model CNN Residual
Total Parameters	364,506	718,586
Trainable Parameters	364,506	718,586
Non-trainable Parameters	0	0
Parameter Size (MB)	1,39	2,74

With a total of 15,000 voice sample files, Figure 10. shows that the CNN DSC model trains 5.12 seconds faster than the CNN Residual model. This is based on tests that compared the performance of the training process time for voice biometrics on the two CNN models. The decreased size of the trainable parameters and parameter size in the CNN DSC model compared to the CNN residual model results in a lower computational load. The findings of this comparison indicate that the training process time performance of CNN DSC is superior to that of CNN Residual, with a time difference of 2.25 seconds. To decrease the computational burden of the training process, one can achieve this by minimizing the number of training parameters and accelerating the training period for the CNN DSC model. This will significantly decrease the substantial computational operational expenses, in contrast to the CNN residual approach. Figure 4.5 displays the results of the performance testing for the training process time of voice biometrics on the CNN DSC and CNN Residual models.

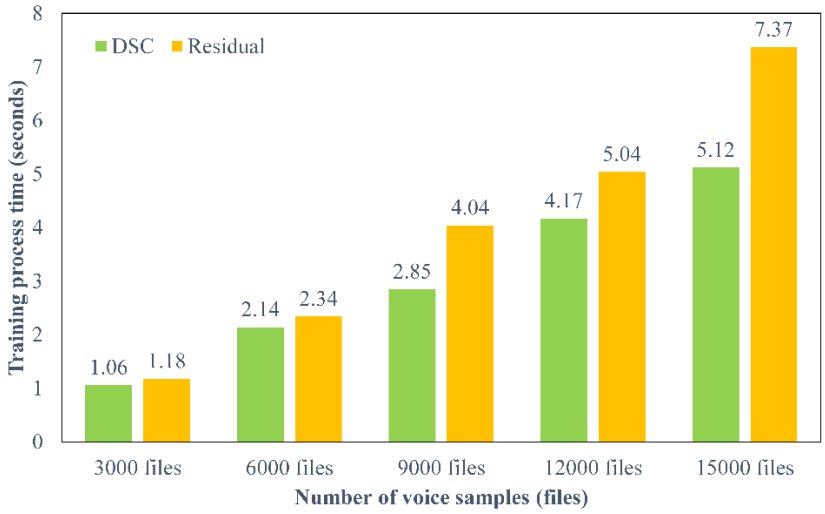


Figure 10. Performance Testing Comparison of Training Process Time on Voice Biometrics using the model of CNN DSC and CNN Residual

4.3. Comparative analysis of speaker recognition performance testing (Who is speaking?) between the CNN DSC and CNN residual models

The purpose of this speaker recognition performance test is to evaluate the performance of speaker recognition utilizing the deep learning algorithms CNN Residual and CNN DSC, employing the DWT-MFCC technique. The speaker recognition performance test yielded accuracy and precision numbers for speaker recognition using the CNN residual and CNN DSC model algorithms. The test was conducted on voice sample files ranging from 3,000 to 15,000 files. The results of the speaker recognition performance analysis utilizing the CNN residual and CNN DSC model methods are presented in Figure 11.

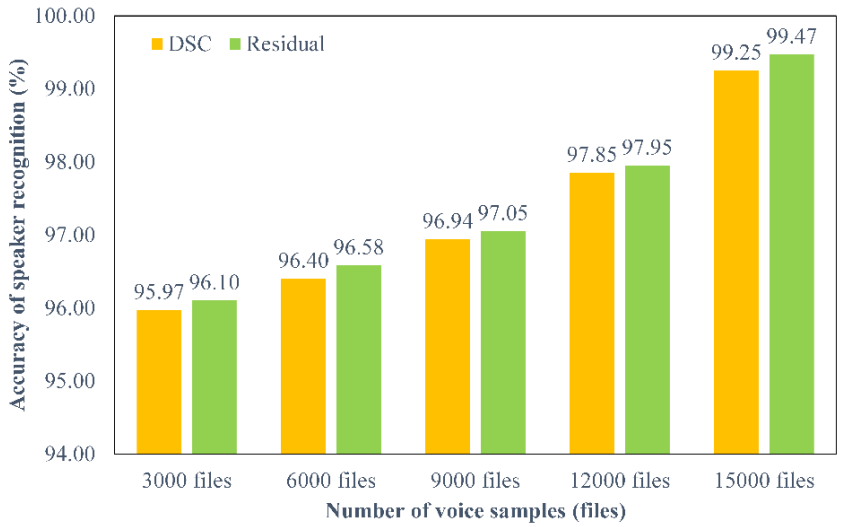


Figure 11. Comparison of Speaker Recognition Accuracy Performance using the model of CNN DSC and CNN Residual

The speaker recognition performance of the CNN residual and CNN DSC models utilizing DWT-MFCC was compared. The CNN residual model achieved the highest accuracy, with a percentage accuracy value of 99.47%. This result was obtained using a total of 15,000 voice sample files. According to the examination of performance data, it was discovered that as the number of voice sample files being processed increases, the accuracy of classification prediction in detecting or identifying the user's voice also increases. To achieve a high level of accuracy in the parameter values during the test assessment (confusion matrix), it is essential to minimize the occurrence of false positives (FP) and false negatives (FN) in order to avoid errors in user voice recognition. High values of the test parameters FP and FN indicate a resemblance between the user's voice and that of other users, resulting in a drop in accuracy. Figure 12 illustrates this.

Based on the results of a comparison of how well the CNN residual model and the CNN DSC model using DWT-MFCC recognized speakers, the CNN residual model got the highest precision percentage value of 99.91%. This result was acquired using a total of 15,000 voice sample files. According to the examination of performance data, it was discovered that as the number of voice sample files executed increases, the accuracy of classification predictions also increases. Figure 12. illustrates this. According to the data analysis of the performance comparison findings mentioned above, the CNN residual model achieves the highest percentage value, indicating the best performance in terms of speaker recognition accuracy and precision. The CNN residual model has superior speaker recognition performance compared to the CNN DSC model. In this scenario, achieving a high accuracy value is contingent upon obtaining a low false positive (FP) value in the evaluation test (confusion

matrix) while simultaneously achieving a high true positive (TP) value. This ensures that the user's voice test is accurately detected as true, without any occurrence of similarity with other user voices. Consequently, when the number of false positives increases, the accuracy value lowers, affecting its reliability.

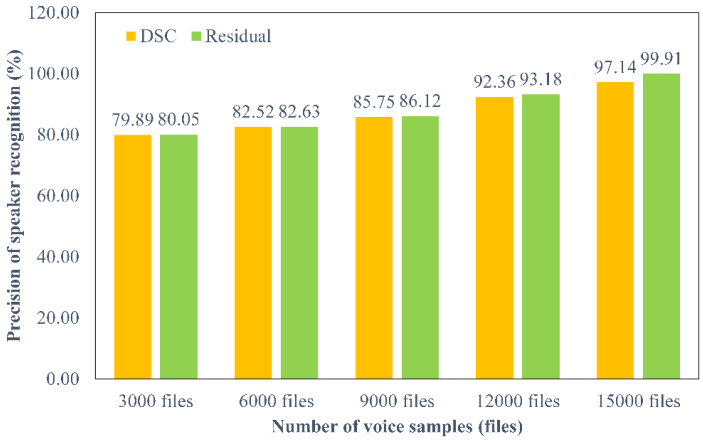


Figure 12. Comparison of Speaker Recognition Precision Performance using the model of CNN DSC and CNN Residual

4.4. Comparative Analysis of Speech Recognition Performance Testing (What was said?) between the CNN Residual Model and CNN DSC

This performance test evaluates the accuracy of speech recognition between the CNN residual algorithm and the CNN DSC model using the DWT-MFCC method. The test is conducted on a set of voice sample files consisting of 3,000, 6,000, 9,000, 12,000, and 15,000 files. This is accomplished by comparing keyword speech or matching keyword voice utterances. This test uses Indonesian speakers' vocalizations of the keyword "Open Access." If the statement is deemed appropriate and accurate (true), it will be approved (accepted). Conversely, if the assertion is incorrect or lacks clarity (false), it is promptly dismissed (rejected). Figure 13. displays the analysis data for voice recognition performance testing with the CNN DSC and CNN residual.

Figure 13. demonstrates the implementation of speech recognition performance testing using the CNN residual and CNN DSC model techniques. A total of 10 VUs pronounced the keyword voice, "Open Access," 20 times to test the voice. The test results indicate that the CNN residual model achieved a speech recognition accuracy of 100%, outperforming the CNN DSC model, which achieved an accuracy of 95%.

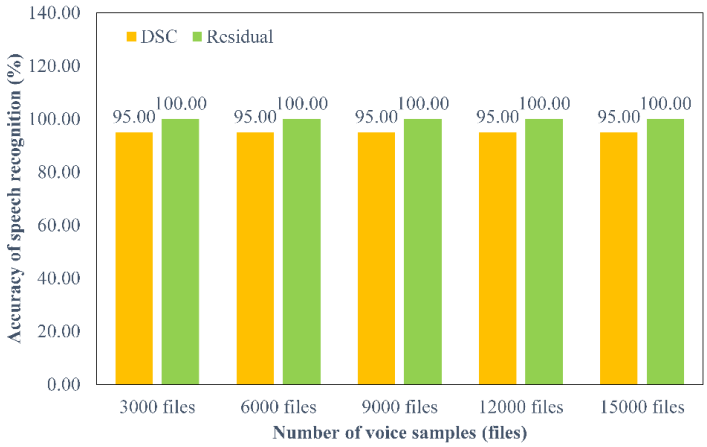


Figure 13. Comparison of Speech Recognition Accuracy Performance in the model of CNN DSC and CNN Residual

The CNN DSC model achieves parameter reduction and minimizes arithmetic operations in convolution processes. This will significantly decrease the expenses associated with operating high-performance computer systems. Therefore, it effectively resolves the issue of excessive computing expenses in the voice classification prediction process as opposed to CNN residual.

4.5. Analysis of response time voice biometrics testing between CNN DSC and CNN Residual models

Response-time testing is conducted to assess the user's speech, specifically in the context of speaker recognition. What is the definition of speech recognition? In order to determine the duration required for detecting user time that has undergone training using the depthwise separable convolution (DSC) CNN and CNN residual models, if the statement is deemed appropriate and accurate (true), it will be approved (accepted). Conversely, if the assertion is incorrect or lacks clarity (false), it is promptly dismissed (rejected). The data analysis for the response time test can be visualized in Figure 14.

Results from the investigation showed that the response time voice biometrics test was done ten times using the CNN DSC and CNN residual algorithm models. Each test utilized 15,000 voice sample files. The acquired results are a comparison of the data on the fastest response time in the two CNN algorithm models.

Based on the test findings, it can be inferred that the CNN DSC model exhibits the shortest test reaction time in comparison to the CNN residual. The CNN DSC model in this scenario offers the benefit of efficient computing, with a total of 364,506 parameters and a parameter size of 1.39 MB. In comparison, the CNN residual has a larger total of 718,586 parameters and a parameter size of 2.74 MB.

The biometrics voice response time test graph revealed that the CNN DSC model had a faster response time, explicitly ranging from 1.54 to 1.82 seconds. Conversely, the CNN residual model exhibited a more significant response time, explicitly ranging from 2.02 to 3.32 seconds.

Based on the test above findings, it can be inferred that the CNN residual model exhibits superiority over the CNN DSC model. Specifically, it enhances validation performance, accuracy in voice biometrics training, precision in speaker detection, and accuracy in speech recognition. The CNN residual model simplifies the training and validation procedures and improves accuracy in prediction classification.

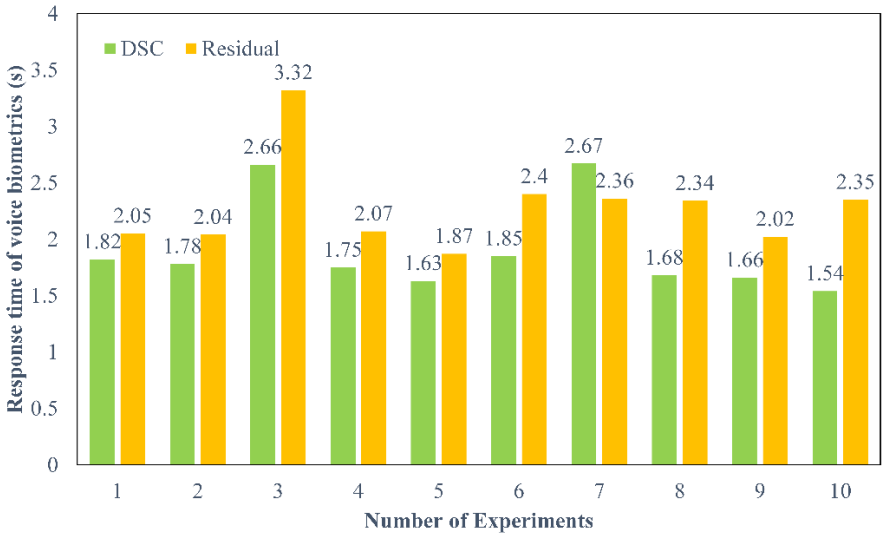


Figure 14. Comparison of the response time of voice biometrics in the CNN DSC and CNN Residual model algorithms

On the other hand, the CNN DSC model has the advantage of having significantly smaller values for both the total parameter and parameter size, which are around half the size of those in the CNN residual model. This impacts the duration of the training process, the speed at which speech biometrics may be identified, and the computational workload required. The CNN DSC model provides a more expedited training method compared to the CNN residual model. What this study aims to do is develop advanced deep learning algorithms for CNN residual and CNN DSC using a mixed method that combines discrete wavelet transform (DWT) and mel-frequency cepstral coefficients (MFCC). The algorithms will be developed to precisely, rapidly, and effectively detect and verify user voice biometrics while maintaining a high level of accuracy for system entry. Guarantee the safeguarding and maintenance of user identity confidentiality.

Based on the aforementioned test findings, it can be inferred that the CNN residual model outperforms the CNN DSC model in several aspects. Specifically, it enhances validation performance and accuracy in voice biometrics training, as well as accuracy and precision in speaker recognition. The CNN residual model simplifies the training and validation processes and improves prediction and classification accuracy. Nevertheless, this CNN residual model suffers from the drawback of having large values for both the total parameter and parameter size. When the CNN residual model is used in the training process, it may result in a longer training period and an increased computing burden compared to the CNN DSC model.

However, the CNN DSC model has the advantage of having fewer total parameter and parameter size values, which are nearly twice as high as the CNN Residual model. The CNN DSC model exhibits shorter training process times and reduced computing burdens compared to the CNN residual model. The reason for this is because the CNN DSC model is capable of reducing the number of parameters and minimizing arithmetic operations during convolution processes. As a result, the exorbitant operating expenses related to computers will be reduced. Therefore, it effectively resolves the issue of excessive computing expenses in voice classification prediction, in contrast to CNN residual.

The CNN DSC model has a drawback in that its accuracy performance is good, but it is somewhat inferior compared to the accuracy performance of the CNN Residual model. Nevertheless, the CNN DSC model outperforms the CNN Standard model in terms of accuracy and performance.

This research aims to develop deep learning algorithms for CNN Residual and CNN DSC using hybrid DWT-MFCC. The algorithms will accurately, quickly, and efficiently identify and verify user voice biometrics. The research will also focus on maintaining high accuracy for security system access and user identity privacy. Based on the findings of the study above, it is advisable to employ the CNN DSC model to identify and verify user voice biometrics. The CNN DSC model has several advantages, including excellent accuracy, parameter reduction, and decreased arithmetic operations in convolution processes. This will significantly reduce computational expenses and improve the speed of the user's speech categorization process. As a result, it effectively resolves the issue of excessive computational costs in voice classification prediction compared to CNN residual.

5. Conclusions

The CNN residual model's voice biometrics testing reveals a training accuracy of 98.6345% and a validation performance of A large number of CNN residual training parameters, namely 718,586 parameters, leads to a lengthy training process lasting 7.37 seconds and a response time of 2.35 seconds. As a result, the amount of computing labor becomes greater. The CNN residual has a more significant number of parameters due to the presence of eight more shortcut connections in its design. This shortcut link has a positive effect on enhancing classification accuracy and performance. The CNN residual speaker recognition system achieves outstanding accuracy and precision, with performance values of 99.47% and 99.91%, respectively. The accuracy of the voice

recognition of CNN Residual is remarkable, and it has achieved a flawless score of 100%. The restricted user sample size, consisting of just 10 individuals, impacts the accuracy of user categorization predictions, leading to a higher level of performance evaluation. However, the upcoming difficulty is augmenting the quantity of voice users to a total of 30 individuals.

The CNN DSC model was subjected to voice biometrics testing, resulting in training accuracy and validation performance ratings of 98.3542%, which were quite outstanding. The CNN DSC decreased the training parameter count to 364,506 in comparison to the CNN Residual. This resulted in a decrease in the computational workload and a reduction in the training process duration to 5.12 seconds, with the fastest reaction time being 1.54 seconds. Recent research done by CNN DSC on speaker recognition has shown outstanding performance, achieving accuracy and precision rates of 99.25% and 97.14%, respectively. Decreasing the training parameters in CNN DSC will lead to reduced memory needs, decreased computational burden, and improved training time efficiency. As a result, it effectively solves the problem of expensive computational costs. The accuracy of the voice recognition achieved on CNN DSC demonstrates an impressive degree of performance, reaching 95%. The CNN DSC's exceptional precision in forecasting voice categorization will enhance user identity protection in banking transactions. According to the investigation, the test results presented above demonstrate outstanding performance. The CNN DSC approach efficiently decreases the number of parameters and minimizes arithmetic calculations during convolution procedures. This has the potential to reduce the amount of computing work required and improve the security of the user's identification system, leading to more accurate and efficient financial transactions. Therefore, our CNN DSC model may effectively solve the problem of high computational costs. According to the study findings described earlier, it is recommended that the CNN DSC model be adopted for the task of recognizing and confirming user voice biometrics. The CNN DSC model has several benefits, such as exceptional accuracy, reduced parameters, and lowered arithmetic operations during convolution procedures. Implementing this will significantly reduce computing costs and speed up the user's voice classification process. Consequently, it efficiently addresses the problem of high computational expenses in audio classification prediction as compared to CNN residual. Future research may include developing and deploying a real-time voice biometrics framework system to instantly identify and authenticate user voices for safe access to Internet banking services.

References

Alsobhani, A., Alabboodi, H. M. A., & Mahdi, H. (2021). Speech Recognition using Convolution Deep Neural Networks. *Journal of Physics: Conference Series*, *1973*(1). https://doi.org/10.1088/1742-6596/1973/1/012166

Amjad Hassan Khan, & P. S. Aithal. (2022). Voice Biometric Systems for User Identification and Authentication – A Literature Review. *International Journal of Applied Engineering and Management Letters (IJAEML) A Refereed International Journal of Srinivas University*, *6*(1), 2581–7000.

Andra, M. B., & Usagawa, T. (2021). Improved Transcription and Speaker Identification System for Concurrent Speech in Bahasa Indonesia Using Recurrent Neural Network. *IEEE Access*, *9*, 70758–70774. https://doi.org/10.1109/ACCESS.2021.3077441

Arora, S., & Bhatia, M. P. S. (2022). Challenges and opportunities in biometric security: A survey. *Information Security Journal: A Global Perspective*, *31*(1), 28–48. https://doi.org/10.1080/19393555.2021.1873464

Batista, G. C., Oliveira, D. L., Saotome, O., & Silva, W. L. S. (2020). A low-power asynchronous hardware implementation of a novel SVM classifier, with an application in a speech recognition system. *Microelectronics Journal*, *105*, 104907. https://doi.org/https://doi.org/10.1016/j.mejo.2020.104907

Chai, L., Du, J., Liu, Q.-F., & Lee, C.-H. (2021). A Cross-Entropy-Guided Measure (CEGM) for Assessing Speech Recognition Performance and Optimizing DNN-Based Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 106–117. https://doi.org/10.1109/TASLP.2020.3036783

Chowdhury, A., & Ross, A. (2020). Fusing MFCC and LPC Features Using 1D Triplet CNN for Speaker Recognition in Severely Degraded Audio Signals. *IEEE Transactions on Information Forensics and Security*, *15*, 1616–1629. https://doi.org/10.1109/TIFS.2019.2941773

Duraibi, S., Sheldon, F. T., & Alhamdani, W. (2020). Voice Biometric Identity Authentication Model for IoT Devices. *International Journal of Security,*

Privacy and Trust Management, *9*, 1–10. https://doi.org/10.5121/ijsptm.2020.9201

Filho, E. M. D. L., Filho, G. P. P. R., Sousa, R. T. De, & Gonçalves, V. P. (2022). Improving Data Security, Privacy, and Interoperability for the IEEE Biometric Open Protocol Standard. *IEEE Access*, *10*, 26985–27001. https://doi.org/10.1109/ACCESS.2020.3046630

Hao, Q., Wang, F., Ma, X., & Zhang, P. (2021). A Speech Recognition Algorithm of Speaker-Independent Chinese Isolated Words Based on RNN-LSTM and Attention Mechanism. *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 1–4. https://doi.org/10.1109/CISP-BMEI53629.2021.9624368

Hidayat, R., & Winursito, A. (2020). A Modified MFCC for Improved Wavelet-Based Denoising on Robust Speech Recognition. *International Journal of Intelligent Engineering and Systems*, *14*(1), 12–21. https://doi.org/10.22266/IJIES2021.0228.02

Huang, C., Zhu, Z., & Guo, J. (2020). Investigations of HMM-Based Speech Recognition Technology. *2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)*, 74–77. https://doi.org/10.1109/IWECAI50956.2020.00021

Ibrahim, W., Candra, H., & Isyanto, H. (2022). Voice Recognition Security Reliability Analysis Using Deep Learning Convolutional Neural Network Algorithm. *Journal of Electrical Technology UMY*, *6*(1), 1–11. https://doi.org/10.18196/jet.v6i1.14281

Ihsanto, E., Ramli, K., Sudiana, D., & Gunawan, T. S. (2020a). An efficient algorithm for cardiac arrhythmia classification using ensemble of depthwise separable convolutional neural networks. *Applied Sciences (Switzerland)*, *10*(2). https://doi.org/10.3390/app10020483

Ihsanto, E., Ramli, K., Sudiana, D., & Gunawan, T. S. (2020b). Fast and accurate algorithm for ECG authentication using residual depthwise separable convolutional neural networks. *Applied Sciences (Switzerland)*, *10*(9). https://doi.org/10.3390/app10093304

Isyanto, H., Arifin, A. S., & Suryanegara, M. (2022). Voice Biometrics for Indonesian Language Users using Algorithm of Deep Learning CNN Residual and Hybrid of DWT-MFCC Extraction Features. *International Journal of Advanced Computer Science and Applications*, *13*(5), 622–634. https://doi.org/10.14569/IJACSA.2022.0130574

Jolad, B., & Khanai, R. (2022). *ANNs for Automatic Speech Recognition—A Survey* (pp. 35–48). https://doi.org/10.1007/978-981-16-2126-0_4

Jung, S.-Y., Liao, C.-H., Wu, Y.-S., Yuan, S.-M., & Sun, C.-T. (2021). Efficiently Classifying Lung Sounds through Depthwise Separable CNN Models with Fused STFT and MFCC Features. *Diagnostics (Basel, Switzerland)*, *11*(4). https://doi.org/10.3390/diagnostics11040732

Kanervisto, A., Hautamäki, V., Kinnunen, T., & Yamagishi, J. (2022). Optimizing Tandem Speaker Verification and Anti-Spoofing Systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 477–488. https://doi.org/10.1109/TASLP.2021.3138681

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(12), 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827

Liu, L. (2022). The New Approach Research on Singing Voice Detection Algorithm Based on Enhanced Reconstruction Residual Network. *Journal of Mathematics*, *2022*, 7987592. https://doi.org/10.1155/2022/7987592

Lu, G., Zhang, W., & Wang, Z. (2022). Optimizing Depthwise Separable Convolution Operations on GPUs. *IEEE Transactions on Parallel and Distributed Systems*, *33*(1), 70–87. https://doi.org/10.1109/TPDS.2021.3084813

M S, A., & P S, S. (2021). Classification of Pitch and Gender of Speakers for Forensic Speaker Recognition from Disguised Voices Using Novel Features Learned by Deep Convolutional Neural Networks. *Traitement Du Signal*, *38*, 221–230. https://doi.org/10.18280/ts.380124

Malik, R. A., Setianingsih, C., & Nasrun, M. (2020). Speaker Recognition for Device Controlling using MFCC and GMM Algorithm. *2020 2nd International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)*, 1–6. https://doi.org/10.1109/ICECIE50279.2020.9309603

Moreno, L. C., & Lopes, P. B. (2018). Voice Biometrics Based on Pitch Replication. *International Journal for Innovation Education and Research*, *6*(10), 351–358. https://doi.org/10.31686/ijier.vol6.iss10.1201

Nainan, S., & Kulkarni, V. (2021). Enhancement in speaker recognition for optimized speech features using GMM, SVM and 1-D CNN. *International Journal of Speech Technology*, *24*(4), 809–822. https://doi.org/10.1007/s10772-020-09771-2

Nayana, P. K., Mathew, D., & Thomas, A. (2017). Comparison of Text Independent Speaker Identification Systems using GMM and i-Vector Methods. *Procedia Computer Science*, *115*, 47–54. https://doi.org/https://doi.org/10.1016/j.procs.2017.09.075

Pawade, D., Sakhapara, A., Ashtekar, R., Bakhai, D., & Tyagi, S. (2022). Voice Based Authentication Using Mel-Frequency Cepstral Coefficients and Gaussian Mixture Model. *2022 IEEE Bombay Section Signature Conference (IBSSC)*, 1–6. https://doi.org/10.1109/IBSSC56953.2022.10037421

Ping, L. (2021). English Speech Recognition Method Based on HMM Technology. *2021 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, 646–649. https://doi.org/10.1109/ICITBS53129.2021.00164

Pyykkönen, P., Mimilakis, S. I., Drossos, K., & Virtanen, T. (2020). Depthwise Separable Convolutions Versus Recurrent Neural Networks for Monaural Singing Voice Separation. *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 1–6. https://doi.org/10.1109/MMSP48831.2020.9287169

Quang, C. T., Nguyen, Q. M., Phuong, P. N., & Do, Q. T. (2021). Improving Speaker Verification in Noisy Environment Using DNN Classifier. *2021 RIVF International Conference on Computing and Communication Technologies (RIVF)*, 1–5. https://doi.org/10.1109/RIVF51545.2021.9642074

Rui, Z., & Yan, Z. (2019). A Survey on Biometric Authentication: Toward Secure and Privacy-Preserving Identification. *IEEE Access*, *7*, 5994–6009.

https://doi.org/10.1109/ACCESS.2018.2889996

Sarkar, A., & Singh, B. K. (2020). A review on performance,security and various biometric template protection schemes for biometric authentication systems. *Multimedia Tools and Applications*, *79*(37), 27721–27776. https://doi.org/10.1007/s11042-020-09197-7

Sen, N., Sahidullah, M., Patil, H., Mandal, S., Rao, K., & Basu, T. (2021). Utterance partitioning for speaker recognition: an experimental review and analysis with new findings under GMM-SVM framework. *International Journal of Speech Technology*, *Article in*. https://doi.org/10.1007/s10772-021-09862-8

Shan, W., Yang, M., Wang, T., Lu, Y., Cai, H., Zhu, L., Xu, J., Wu, C., Shi, L., & Yang, J. (2021). A 510-nW Wake-Up Keyword-Spotting Chip Using Serial-FFT-Based MFCC and Binarized Depthwise Separable CNN in 28-nm CMOS. *IEEE Journal of Solid-State Circuits*, *56*(1), 151–164. https://doi.org/10.1109/JSSC.2020.3029097

Shan, W., Yang, M., Xu, J., Lu, Y., Zhang, S., Wang, T., Yang, J., Shi, L., & Seok, M. (2020). 14.1 A 510nW 0.41V Low-Memory Low-Computation Keyword-Spotting Chip Using Serial FFT-Based MFCC and Binarized Depthwise Separable Convolutional Neural Network in 28nm CMOS. *2020 IEEE International Solid-State Circuits Conference - (ISSCC)*, 230–232. https://doi.org/10.1109/ISSCC19947.2020.9063000

Sholokhov, A., Kinnunen, T., Vestman, V., & Lee, K. A. (2020). Voice biometrics security: Extrapolating false alarm rate via hierarchical Bayesian modeling of speaker verification scores. *Computer Speech & Language*, *60*, 101024. https://doi.org/https://doi.org/10.1016/j.csl.2019.101024

Singh, G., Bhardwaj, G., Singh, S. V., & Garg, V. (2021). *Biometric Identification System: Security and Privacy Concern BT - Artificial Intelligence for a Sustainable Industry 4.0* (S. Awasthi, C. M. Travieso-González, G. Sanyal, & D. Kumar Singh (eds.); pp. 245–264). Springer International Publishing. https://doi.org/10.1007/978-3-030-77070-9_15

Singh, M. K. (2023). A text independent speaker identification system using ANN, RNN, and CNN classification technique. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-023-17573-2

Singla, D., & Verma, N. (2023). Machine and Deep learning in Biometric Authentication: A Review. *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, 22–26. https://doi.org/10.1109/InCACCT57535.2023.10141692

Taye, M. M. (2023). Theoretical understanding of convolutional neural network: concepts, architectures, applications, future directions. *Computation*, *11*(3), 52. https://doi.org/https://doi.org/10.3390/computation11030052

Tyagi, A., Ipsita, Simon, R., & khatri, S. K. (2019). Security Enhancement through IRIS and Biometric Recognition in ATM. *2019 4th International Conference on Information Systems and Computer Networks (ISCON)*, 51–54. https://doi.org/10.1109/ISCON47742.2019.9036156

Wei, Y. (2020). Adaptive Speaker Recognition Based on Hidden Markov Model Parameter Optimization. *IEEE Access*, *8*, 34942–34948. https://doi.org/10.1109/ACCESS.2020.2972511

Wells, A., & Usman, A. B. (2023). Trust and Voice Biometrics Authentication for Internet of Things. *International Journal of Information Security and Privacy*, *17*(1), 1–28. https://doi.org/10.4018/IJISP.322102

Yusuf, N., Marafa, K. A., Shehu, K. L., Mamman, H., & Maidawa, M. (2020). A survey of biometric approaches of authentication. *International Journal of Advanced Computer Research*, *10*(47), 96–104. https://doi.org/10.19101/ijacr.2019.940152

小川充洋. (2021). Parkinson's disease classification by residual network type 1-d CNN using vocal datasets. 生体医工学, *Annual59*(Abstract), 570. https://doi.org/10.11239/jsmbe.Annual59.570