Journal of Electrical Technology UMY (JET-UMY), Vol. 6, No. 1, June 2022
ISSN 2550-1186 e-ISSN 2580-6823

Voice Recognition Security Reliability Analysis Using Deep Learning Convolutional Neural Network Algorithm

Wahyu Ibrahim*[1], Henry Candra[1], Haris Isyanto[2]

[1]Department of Electrical Engineering, Universitas Trisakti
Jl. Kyai Tapa No.1, Kota Jakarta Barat, Daerah Khusus Ibukota Jakarta 11440, Indonesia
[2]Department of Electrical Engineering, Universitas Muhammadiyah Jakarta
Jl Cempaka Putih Tengah 27, Jakarta, Indonesia
*Corresponding author, e-mail: wahyu162012000002@std.trisakti.ac.id

Abstract – *This study discusses the reliability analysis of voice recognition security using the deep learning convolutional neural network (CNN) algorithm. The CNN algorithm has learning advantages in that it is safer, faster, and more accurate. CNN also can solve user identification problems in large amounts of data. The measured voice input is ten types of user's voice with the number of iterations of 6000, 12000, and 15000 sound files. Furthermore, voice extraction features are performed to recognize conversations and retain information that is very much needed. After that, the voice file iteration data is trained to register the user's voice so that a trained model is obtained. These results measure performance (confusion matrix) to analyze the actual value compared to the predicted value in the CNN algorithm. The results obtained are that the best accuracy is obtained at 15000 sound file iterations, 96.87%, 12000 sound file iterations get 96.30%, and 6000 sound file iterations get 95.77%. CNN's performance data shows that 15000 iterations of voice files produce high accuracy. Voice recognition security helps provide high security and maintain the privacy of one's identity.*

Keywords*: voice recognition, convolutional neural network, confusion matrix, accuracy*

I. Introduction

In cyberspace, there are identity theft and data fraud crimes, which can be a new threat for everyone. In terms of accessing information and data identification, a person is very important. So far, the methods used are still common, namely PIN codes, magnetic cards, and passwords. This has weaknesses, for example, misused, damaged, forgotten, lost, stolen, hacked, and counterfeited cards. To reduce this problem, then developed a method of identification for a person [1].

Identification recognition technology is the best solution for maintaining privacy, security, reliability, and speed in processing authentication and individual identification by using a person's biological characteristics. Recognition in a person has a unique characteristic on a person's body, which cannot be imitated. Recognition methods are much more difficult to hack, reconstruct, and fake [1], [2].

Identification recognition has two characteristics: behavioural characteristics and the last is physiological characteristics to know identity authentication. Physiological characteristics have a relationship with the pattern or composition of the human body, such as the pattern of veins, eyes (iris and retina), DNA, hand shape, face shape, fingers, and fingerprints. Behavioural characteristics are related to a unique pattern that can be seen through gait, heart rate, signature, keystroke dynamics, sound, and action [3].

The deep learning part of artificial intelligence is the development of a neural network to provide accuracy in tasks, namely speech recognition and object detection. Deep learning can represent data such as video, text, images, and sound without automatically introducing rules or knowledge of one's domain [4]–[6].

Convolutional neural network (CNN) is a method that can help solve the problem of large amounts of data, and the CNN method can help solve the demands of user identification problems that work

Manuscript received March 2022, revised May 2022

more securely, quickly and accurately [7]–[9]. The results of the convolutional neural network algorithm method measure performance on user voice data in the form of accuracy parameters.

Mel-Frequency Cepstral Coefficients (MFCC) is the best method with high accuracy in identifying and extracting spoken voices to apply the functional principles of the ear artificially. The MFCC can recognize conversations to retain much-needed information, and remove unnecessary noise [18].

A confusion matrix is a method for measuring the performance of an algorithm to analyze the actual value or the actual value to be compared with the predicted value. In this paper, we measure voice recognition voice input data to get a value for generating accuracy parameters using the CNN algorithm [10]–[12].

In this study, choosing the voice recognition method because it has the advantage of being able to authenticate voice, the implementation of voice recognition costs is lower than other recognition because it does not require special devices, such as retina scanners or fingerprint readers, has protection from fraud/fraud, has security high standards, and maintain the privacy of personal identity. The identification recognition method is easy to operate and accurate in individual identification [1], [13]–[16].

In previous writing [17] measured 10 and 20 speaker inputs with a number of sound samples of 320 files and 640 sound files by producing algorithm accuracy ANN-MFCC = 85.3%, SVM-MFCC 64.4%, ANN-LPC 80%, SVM-LPC 71.6%, then at the time of writing [9] measuring five speakers on real-world sound samples resulted in 80% accuracy of the DNN-MFCC algorithm, 90% CNN-MFCC.

Contribution of this research, we propose voice recognition with algorithm of deep learning convolutional neural network (CNN) and MFCC feature extraction. The CNN model has a high accuracy performance that can solve problems with a large number of voice file data and can solve complex data compared to machine learning methods with low accuracy. In addition, other deep learning methods such as ANN and DNN have limitations in data computing capabilities. Furthermore, the MFCC feature extraction method has the advantage of high accuracy in performing voice extraction compared to other feature extraction methods such as LPC, and ZCR. The expected test results of accuracy are >90%. And this research is expected to help provide high security and maintain the privacy of one's identity.

II. Research Method

In this study, the voice recognition system process consists of collecting user voice data composed of 10 types of voices and segmenting 6000, 12000, and 15000 iterations of sound files. The voice collected is carried out feature extraction which aims to retain voice information and dispose of the rest. That is not needed, then the voice is prepared for the voice data training stage using the convolutional neural network algorithm with iterating over the number of voice files so that the user's voice can be registered or labelled. After the training stage is carried out, a trained model is obtained where user voice data has been registered and already stored into the database, the next step is authentication/verification, which is the stage of the user's voice being tested with two processes, namely speaker recognition and speech recognition. Match sound. Where the voice is matched with the knowledge of the trained convolutional neural network (CNN) model that is stored in the database, the results were obtained in the form of validated voice and unfavourable voice in the authentication process (speaker recognition), and keyword verification (speech recognition). The next step is to measure the convolutional neural network algorithm's performance using a confusion matrix that aims to measure the actual value and the predicted value in the user's voice to analyze whether the voice is similar. The results obtained in the confusion matrix performance measurement calculate the accuracy value for the number of iterations of 6000, 12000, and 15000 sound files. The following flowchart of the voice recognition system can be seen in Figure 1.

The next step is to collect user voice input data and collect ten types of user voices. The user's voice samples are taken as many as 6000, 12000, and 15000 sound file samples, then the process of collecting user voice data uses an average sample of 16000 Hz with a microphone device. So that the expected results in collecting user voice data can be processed for extraction features to be able to recognize conversations to retain information, after the feature extraction process is carried out, it can be trained using a convolutional neural network algorithm model to be able to label the user's voice. The following data collection can be seen in Table I.

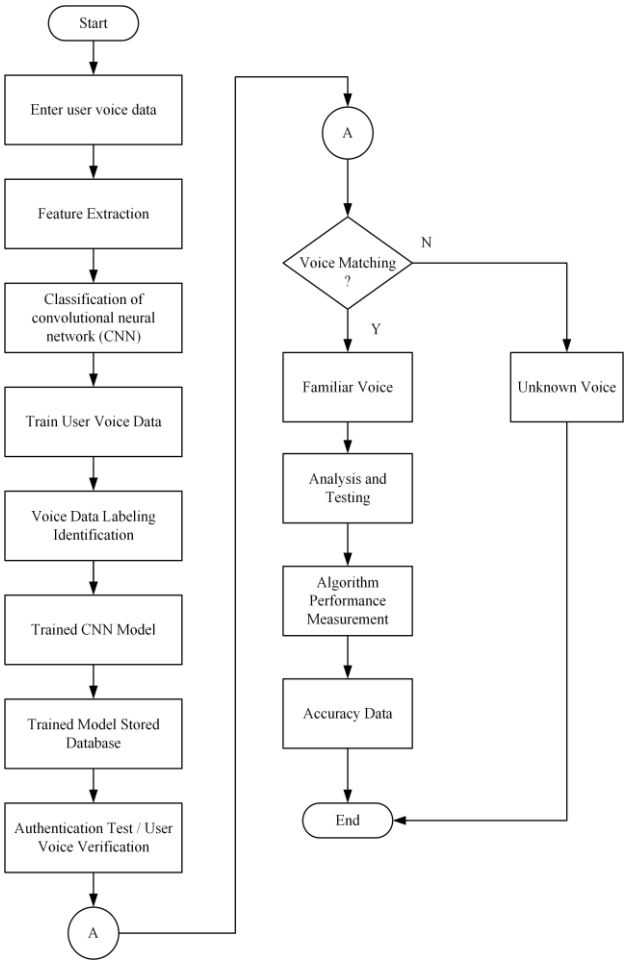
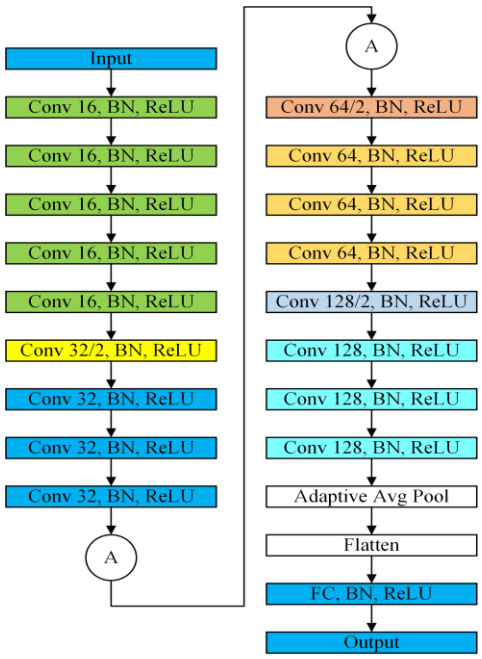


Fig 1. System flowchart of voice recognition

TABLE I
VOICE DATA COLLECTION

Voice Data	Number of Sound Samples (Files)			Sample rate (Hz)
VR_0	600	1200	1500	16000
VR_1	600	1200	1500	16000
VR_2	600	1200	1500	16000
VR_3	600	1200	1500	16000
VR_4	600	1200	1500	16000
VR_5	600	1200	1500	16000
VR_6	600	1200	1500	16000
VR_7	600	1200	1500	16000
VR_8	600	1200	1500	16000
VR_9	600	1200	1500	16000
Total sample	**6000**	**12000**	**15000**	

The next stage in the speaker recognition training process requires the architecture of the convolutional neural network algorithm consisting of input in the form of voice input data with the number of sound files. Convolutional layer with the number of convolution filters/kernels 16, 32, 64, and 128, batch normalization which serves to speed up the training process on voice input data, ReLu is a layer at the network layer to activate the activation function, then the adaptive average pool function to calculate the required kernel. Required to produce an output of the given dimensions. The next stage is the flatten function to convert the data into a 1-dimensional array / one line, and the next stage is connected to a fully connected layer, a full layer process on input data, batch normalization, ReLU, the resulting output is valid and invalid data. The total training parameters on the CNN architecture are 707,386. The following CNN architecture can be seen in Figure 2.



Total training parameters 707,386

Fig 2. CNN Architecture

After knowing the convolutional neural network architecture, the training and authentication stages are carried out on speaker recognition. The training process is training the user's voice data (speakers) to label voices with a convolutional neural network algorithm. At the authentication stage, test the input speaker recognition so that it can be recognized or not through the trained knowledge of the CNN model.

The user's voice data is entered into the convolutional neural network algorithm model at the training stage. Training the data required variations in the total number of sound samples of 6000, 12000, and 15000 sound files. The validation test ratio during training is 10% and requires 40 epochs. The training process aims to label the user's voice (speaker) valid and invalid. The results obtained during training user voice data are in the

form of trained models. The CNN data that has been trained (trained) is stored in the voice recognition database, the following CNN model training data can be seen in Table II.

TABLE II
CNN MODEL TRAINING

Number of Sound Files	Validation Ratio	Epoch
6000		
12000	10%	40
15000		

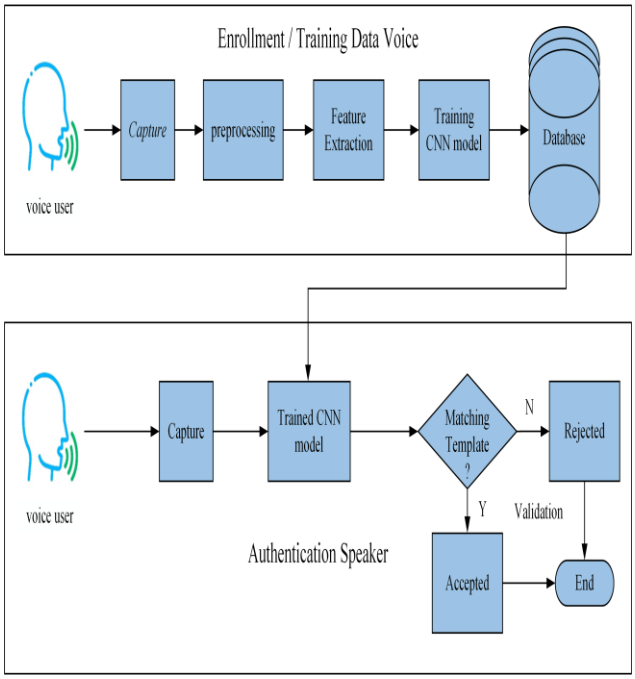


Fig 3. Block Diagram of Speaker Recognition

Next is the speaker recognition stage. There are two processing steps in registration/training data and speaker authentication. In the registration process, where the user's voice is captured (capture) in the form of the voice of who is speaking (speaker), the next stage of the voice is preprocessing to convert the voice signal into a digital signal. Then the voice is extracted using MFCC, which aims to recognize the conversation to retain information and discard the rest that happens to the sound signal. After the feature extraction is carried out, the user voice data training process is carried out with ten types of voices. The voice is trained with variations in the number of voice samples of 6000, 12000, and 15000 voice files to label the user's voice to be valid or invalid, after training the voice sample data, at the training stage will get the results in the form of a trained model. Voices trained (trained) will be stored in the database. Then in the speaker authentication process, at this stage it is recognizing the voice of who is speaking (speaker). This process consists of the user's voice being captured, and then voice matching is done where the matching is on the knowledge of the trained model. So that later votes can be accepted and rejected at validation. The following are the speaker recognition stages, shown in Figure 3.

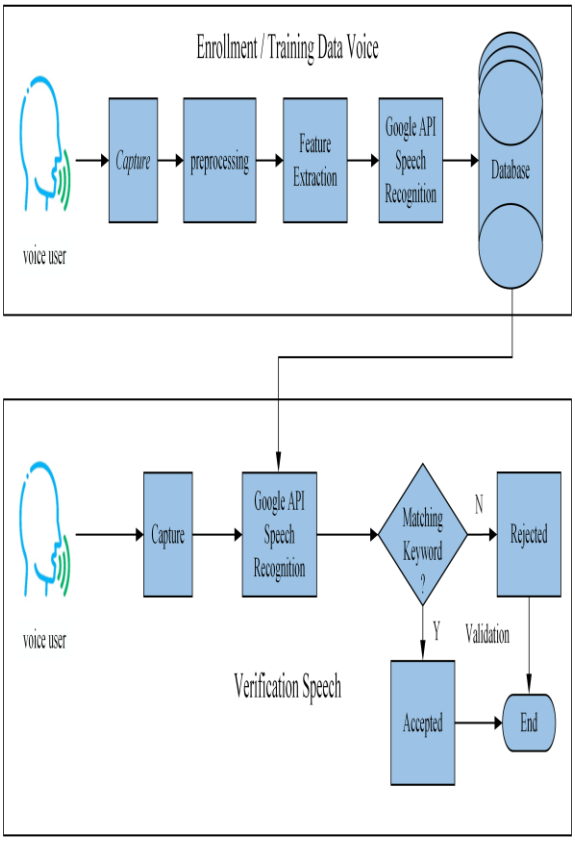


Fig 4. Block Diagram of Speech Recognition

Furthermore, there are two stages of the process at the speech recognition stage, namely registration/training of data and verification of voice content (speech). In the registration process, where the user's voice is captured (capture) in the form of who is speaking (speaker), the next stage is the sound is preprocessed to convert the voice signal into a digital signal. The voice is extracted using MFCC, which aims to recognize the conversation to retain information and remove the rest in the sound signal. After the feature extraction is carried out, the keyword registration process is carried out to open the security of the voice recognition system where keywords are registered using the Google API

speech recognition. After being registered, the keywords are stored in the database. Then in the voice content verification process for keyword verification (speech recognition). This process is carried out by capturing sound. Then the keywords are spoken by the user later from the knowledge of the Google API speech recognition in the form of registered keywords. Then the keywords entered by the user are matched with keywords so that the keywords can be accepted and rejected at the time of validation. The following stages of speech recognition can be seen in Figure 4.

After knowing the block diagram of voice recognition, the next step is testing the performance of the convolutional neural network (CNN) algorithm using a confusion matrix. Where to calculate the algorithm's performance on the number of user voice files. So that the confusion matrix is to analyze the actual value or the actual value compared to the predicted value to get an evaluation matrix, namely measuring accuracy. At the stage of testing the data to measure the algorithm's performance by 10% of the total sound sample file. The following data for measuring the performance of the CNN algorithm can be seen in Table III.

TABLE III
CNN PERFORMANCE MEASUREMENT ALGORITHM

Number of Sound Files	Measurement Test	Sound File Measurement
6000	10%	600
12000		1200
15000		1500

III. Result and Discussion

In this study, analysis of training test data was carried out with the number of iterations of 6000, 12000, and 15000 sound files where to find out the best parameter comparison in that iteration, then analyze the performance of the speaker recognition algorithm on the convolutional neural network algorithm model by measuring the accuracy value in the comparison of the number of file iterations. Voice, test speech recognition to test the success of pronouncing keywords into the voice recognition system, and testing response time to measure how long it takes so that the system can process the input voice.

III.1. CNN model training test

The iteration of the user's voice input data is tested in this test, namely 6000, 12000, and 15000 sound files. The test is to determine the validation of accuracy and validation of loss in the user voice data training process. The training process takes 40 epochs. Where epoch is the process of training user voice data into the CNN algorithm in 1 round, where the process takes 40 rounds, the following data validation accuracy of the number of sound files can be seen in Table IV.

TABLE IV
SOUND FILE ITERATION TRAINING ACCURACY VALIDATION

Epoch	Accuracy Validation (%)		
	6000	12000	15000
10	80.878	82.5064	97.0001
20	95.5132	95.9139	97.2006
30	95.4269	96.0856	97.2391
32	95.3129	96.6235	96.9815
34	95.3991	96.5502	97.4249
38	95.8852	95.7744	97.3279
40	95.4172	96.561	97.2545

Table IV. shows that the best data in the iteration data training process is the number of user voice files. Where the iteration of 6000 sound files obtained the best training data on epoch 38 with an accuracy validation value of 95,8852%, in iteration 12000 sound files obtained in epoch 32 testing the accuracy validation value got 96,6235%, and in iteration 15000 sound files obtained on epoch 34 with the validation value of accuracy is 97.4249%. So that the results obtained are that the greater the number of sound files, the greater the accuracy of the user's voice data training process, the following graph of training accuracy validation can be seen in Figure 5.

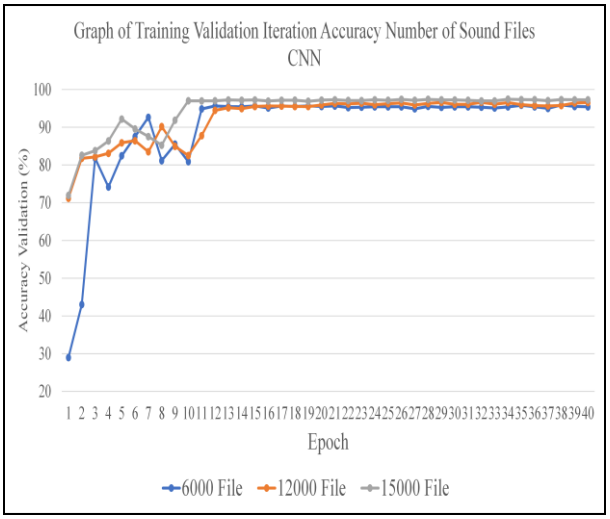


Fig 5. Training Accuracy Validation Graph

Next, measure the loss validation in testing the convolutional neural network training algorithm. The test is carried out with iterations of 6000, 12000, and 15000 sound files with 40 epochs in the training process. The following loss validation data on the training test can be seen in Table V.

TABLE V
SOUND FILE ITERATION TRAINING LOSS VALIDATION

Epoch	Validasi Loss		
	6000	12000	15000
10	0.8452	0.6506	0.1076
20	0.1518	0.1224	0.0978
21	0.1398	0.1163	0.0961
30	0.1461	0.1044	0.0927
34	0.1432	0.1186	0.0891
40	0.1513	0.111	0.0912

In table V. Loss validation values are obtained at 6000 sound file iterations, the best data is received at epoch 21 with a value of 0.1398%, while 12000 user voice files at epoch 30 with a value of 0.1044%, and 15000 sound file iterations are obtained at epoch 34 which is 0.0891%. The results obtained on the loss validation parameter show that the greater the number of iterations of the sound file, the better the loss validation value for iterations of 15000 sound files will be. The graphic data on the sound file iteration training test can be seen in Figure 6.

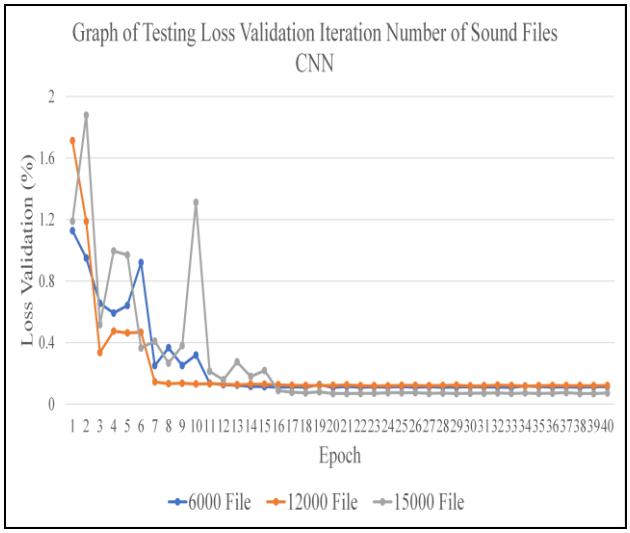


Fig 6. Training Loss Validation Graph

After getting the training parameters of the user's voice data on the convolutional neural network (CNN) algorithm, then the best data is obtained for accuracy validation and loss validation by iterating the number of sound files. So that the data generated is to find out the best results in the training testing process, the following comparison of training tests can be seen in Table VI.

TABLE VI
THE BEST DATA TESTING TRAINING VALIDATION

Accuracy Validation (%)			Loss Validation (%)		
6000	12000	15000	6000	12000	15000
95.88	96.62	97.42	0.1398	0.1044	0.0891

In table VI. It was found that the more sound files in the process of training the user's voice data, the higher the validation accuracy will be. In the iteration test of 15000 sound files, validation accuracy is 97.42%, iteration of 12000 sound files is 96.62%, and iteration of 6000 sound files is 95.88%, so the accuracy validation graph can be seen in Figure 7.

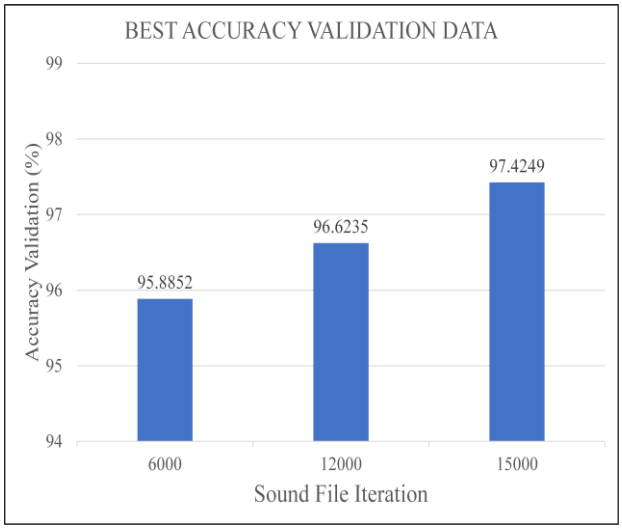


Fig 7. Best Accuracy Validation Graph

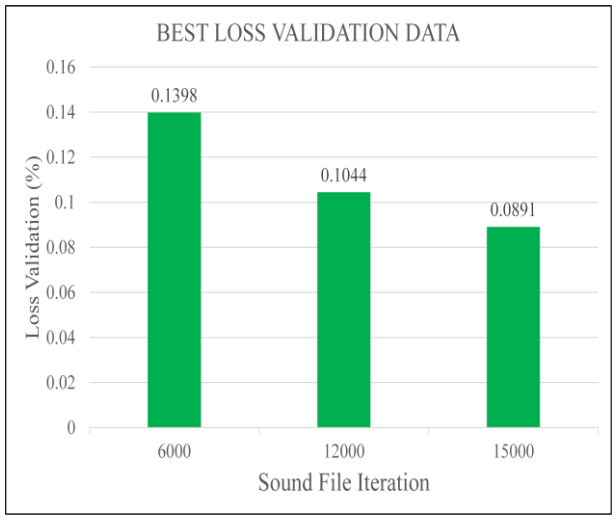


Fig 8. Best Loss Validation Graph

Next is table VI. It was found that the more the number of iterations of the sound file in the testing process, the lowest loss parameter was obtained. In the 15000 iterations of sound files, the loss value is 0.09%, while in the 12000 iterations, the sound file is 0.10%, and the 6000 iteration is 0.14%. The following graph of the best loss validation can be seen in Figure 8.

III.2. CNN speaker recognition performance test

In this study, the performance of the convolutional neural network algorithm was measured. The measured data is in the form of voice input data speaking (speaker). User data is calculated to determine the predicted value and actual value by iterating the number of sound files of 6000, 12000, and 15000. The test measures the input data of 10 types of user voices. Data to measure algorithm performance is 10% of the total file iterations. After knowing the predicted value parameters and the actual values in the form of true positive (TP), false positive (FP), false negative (FN), and true negative (TN) to show similarities to other users' voices, later votes will be used to calculate CNN's performance accuracy in iterations. A number of files. The following algorithm performance data can be seen in table VII.

TABLE VII
PERFORMANCE TEST 6000 SOUND FILES

Number of Samples 6000 Sound Files					
	TP	FP	FN	TN	Accuracy (%)
VR0	55	23	5	517	95.33
VR1	46	3	14	537	97.17
VR2	58	7	2	533	98.50
VR3	52	7	8	533	97.50
VR4	32	26	28	514	91.00
VR5	50	14	10	526	96.00
VR6	56	2	4	538	99.00
VR7	51	7	9	533	97.33
VR8	17	22	43	518	89.17
VR9	56	16	4	524	96.67

In table VII. Actual value and predictive value are obtained for each user's voice data. For VR0, VR4, VR5, and VR8 users, they produce very high false positive (FP) and false-negative (FN) values so that there are similarities between these users. In this case, it can reduce the accuracy obtained. Furthermore, VR6 produces the lowest FP and FN values, so VR6 users get the highest accuracy value, which is 99%. The following graph of speaker recognition performance measurement can be seen in Figure 9.

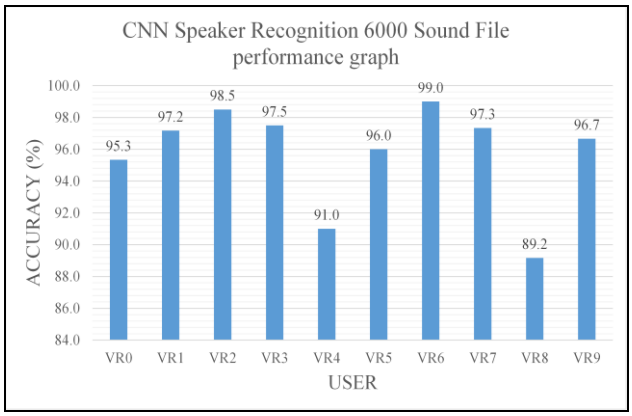


Fig 9. CNN Speaker Recognition 6000 Sound File Performance Graph

Next, measure the iteration performance of 12000 votes using the CNN algorithm. The test is carried out to determine the actual and predicted values to calculate the accuracy value in the 12000 iterations of the number of votes. The following iteration performance data can be seen in Table VIII.

TABLE VIII
PERFORMANCE TEST 12000 SOUND FILES

Number of Samples 12000 Sound Files					
	TP	FP	FN	TN	Accuracy (%)
VR0	159	5	6	430	98.17
VR1	163	36	2	399	93.67
VR2	163	14	2	421	97.33
VR3	149	12	16	423	95.33
VR4	150	2	15	433	97.17
VR5	152	0	13	435	97.83
VR6	159	3	6	432	98.50
VR7	146	6	19	429	95.83
VR8	154	9	11	426	96.67
VR9	143	25	22	410	92.17

In table VIII. Obtained the actual value and the predicted value of the user's voice. The data obtained on the VR1 and VR9 users where the data obtained on the FP and FN parameters are very high. So that the accuracy received decreases compared to other users' voices, VR6 users get the best accuracy compared to other voice users, and the accuracy is 98.50%. The following graph of the performance of the speaker recognition iteration of 12000 sound files can be seen in Figure 10.

Next, measure the iteration performance of 15000 sound files on the CNN algorithm. The measurement to find out the actual value is compared with the predicted value to analyze the similarity that occurs with the voices of other users, after getting that value to calculate accuracy on 15000 iterations of sound files. The following iteration performance data for 15000 sound files can be seen in table IX.

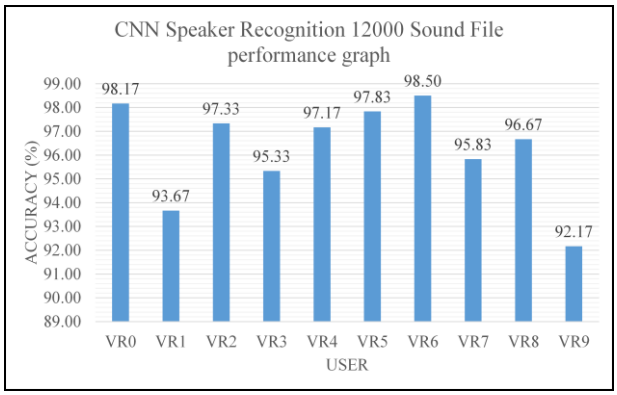


Fig 10. CNN Speaker Recognition 12000 Sound File Performance Graph

TABLE IX
PERFORMANCE TEST 12000 SOUND FILES

Number of Samples 15000 Sound Files					
	TP	FP	FN	TN	Accuracy (%)
VR0	189	4	3	404	98.8
VR1	192	31	0	377	94.8
VR2	192	22	0	386	96.3
VR3	175	3	17	405	96.7
VR4	183	8	9	400	97.2
VR5	191	9	1	399	98.3
VR6	186	6	6	402	98.0
VR7	163	1	29	407	95.0
VR8	174	2	18	406	96.7
VR9	181	8	11	400	96.8

In table IX. The speaker recognition performance value is obtained with iterations of 15000 sound files in the form of parameters TP, FP, FN, TN. VR1, VR2, and VR7 users experience a decreased inaccuracy. In this case, the decrease in accuracy parameters is due to the high values of FP and FN. The effect of the high value will be the sound similarity. When the FP and FN values are low, they will produce high accuracy. VR0 users have a high accuracy value of 98.8%. High accuracy is obtained with low FP and FN values. The following graph of the CNN speaker recognition performance iteration 15000 sound files can be seen in Figure 11.

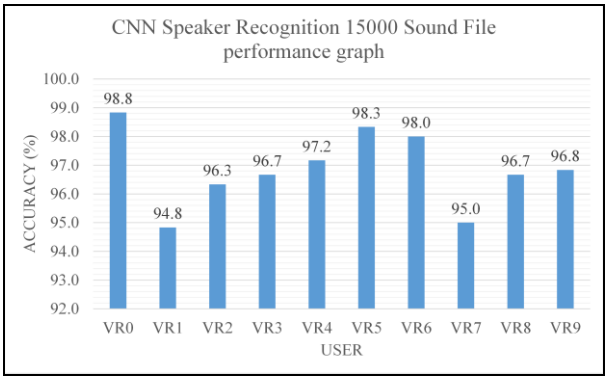


Fig 11. CNN Speaker Recognition 15000 Sound File Performance Graph

After testing the performance on 6000, 12000, and 15000 iterations of sound files, a comparison was made to determine the best number of iterations on the convolutional neural network (CNN). The results were obtained using the best accuracy from iteration testing. The iteration performance comparison data can be seen in Table X.

TABLE X
BEST ITERATION PERFORMANCE COMPARISON DATA

User	Accuracy (%) *Convolutional neural network*		
	6000	12000	15000
VR0	95.33	98.17	98.83
VR1	97.17	93.67	94.83
VR2	98.50	97.33	96.33
VR3	97.50	95.33	96.67
VR4	91.00	97.17	97.17
VR5	96.00	97.83	98.33
VR6	99.00	98.50	98.00
VR7	97.33	95.83	95.00
VR8	89.17	96.67	96.67
VR9	96.67	92.17	96.83
Average	**95.77**	**96.30**	**96.87**

In table X., It is found that the 15000 iterations of the number of sound files got the best results, namely 96.87% compared to the 6000 iterations of 95.77%, and 12000 sound files of 96.30%. So that the more the number of sound files in the test, the accuracy will increase, in this case, the 15000 sound file iteration has increased accuracy compared to other iterations. The following graph of accuracy comparison can be seen in Figure 12.

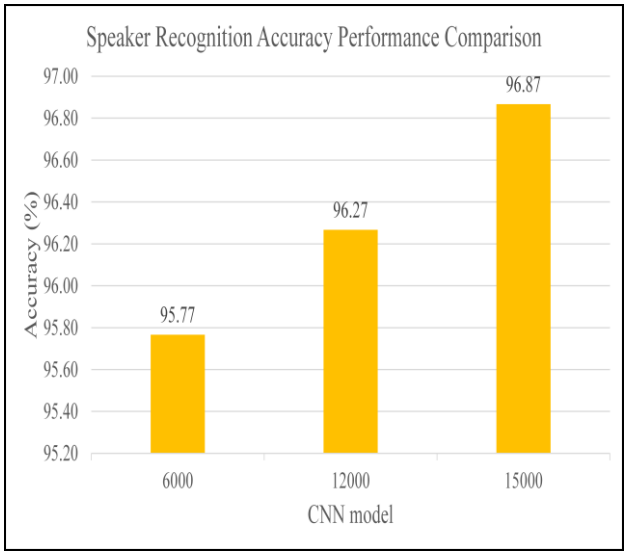


Fig 12. Speaker Recognition Performance Comparison

III.3. Testing the voice recognition system

This study testing authentication of speaker recognition and verification of keywords (speech recognition). Where the test performs, two security passes so that it can open the voice recognition system. When the user's voice is spoken, it will produce two outputs: speaker recognition authentication and speech recognition keyword verification. If you have passed the two security points, the voice recognition system will be accepted in the next step. If not, it will be rejected. The following data for testing the voice recognition system can be seen in table XI.

TABLE XI
TESTING THE VOICE RECOGNITION SYSTEM

Test	Speaker recognition		Speech Recognition		Voice recognition	
	True (T)	False (F)	True (T)	False (F)	Accepted (A)	Rejected (R)
1	T		T		A	
2	T		T		A	
3	T		T		A	
4	T		T		A	
5	T		T		A	
6	T		T		A	
7	T		T		A	
8	T		T		A	
9	T		T		A	
10	T		T		A	
11	T		T		A	
12	T		T		A	
13	T		T		A	
14	T		T		A	
15	T		T		A	
16	T		T		A	
17	T		T		A	
18		F		F		R
19	T		T		A	
20	T		T		A	
Average	**95%**	**5%**	**95%**	**5%**	**95%**	**5%**

In table XI. Conducted 20 experiments to determine the parameters obtained in the voice recognition system. Where the average voice recognition system testing is in the form of threshold parameters, presentation of the success of the speaker recognition system, and speech recognition, the data obtained by testing the user's voice at an average threshold value of 0.77, the presentation of the success of the voice recognition system is 95%. The percentage of failure is 5%.

III.4. Voice recognition time response test

In this paper, we measure the response time of the voice recognition system. In this test, the voice is entered and then calculates the response time of the voice recognition system to be able to enter the system. The following data for testing the time response of the voice recognition system can be seen in table XII.

TABLE XII
VOICE RECOGNITION TIME RESPONSE TEST

Test	**Respon time (s)** ***voice recognition***
1	4.13
2	3.92
3	3.85
4	3.78
5	3.83
6	3.55
7	4.19
8	3.46
9	4.05
10	3.70
Average	**3.85**

In Table XI. The value of the response time of the voice recognition system is obtained. Where is the response time to find out the time needed to open voice recognition security? The response time of the voice recognition system depends on the speed or stability of the internet. In testing, the response time obtained an average of 3.85 seconds with ten trials. The following graph of the response time of the voice recognition system can be seen in Figure 13.

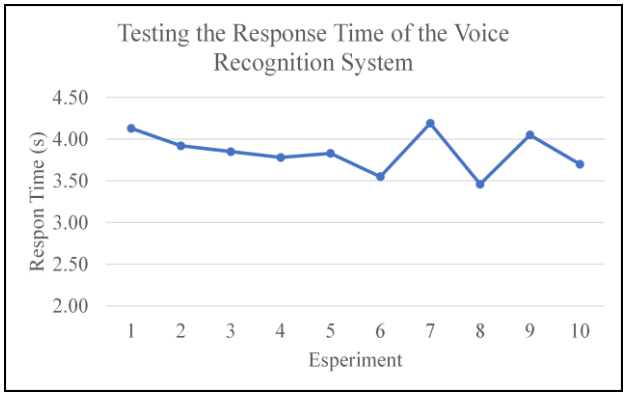


Fig 13. Voice Recognition Response Time Graph

IV. Conclusion

1. The larger the number of sound file iterations will produce training parameters and excellent performance.
2. The results of the training test at iteration 15000 obtained the best accuracy validation value of 97.42%, while the validation loss value was 0.891%
3. The results of testing the performance of the CNN algorithm using a confusion matrix shows that in iterations of 15000 sound files,

the best accuracy is 96.87%. In comparison, for 12000 sound files, the accuracy is 96.30%, and for an iteration of 6000 sound files, accuracy is 95.77%.

4. The successful voice recognition system testing results with 20 system trials obtained a 95% success presentation and 5% failure presentation.
5. The results of the test response time for voice recognition, the average time to read the user's voice takes 3.85 seconds.
6. The voice recognition system is very appropriate to be implemented because it has two securities to open the system, which requires speaker recognition authentication, and voice content verification (speech recognition).

Acknowledgements

The author would like to thank the Department of Electrical Engineering, Faculty of Industrial Technology, Trisakti University for their support.

References

[1] Z. Rui and Z. Yan, "A Survey on Biometric Authentication: Toward Secure and Privacy-Preserving Identification," *IEEE Access*, vol. 7, pp. 5994–6009, 2019, doi: 10.1109/ACCESS.2018.2889996.

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

[3] J. Handa, S. Singh, and S. Saraswat, "Approaches of Behavioural Biometric Traits," in *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2019, pp. 516–521, doi: 10.1109/CONFLUENCE.2019.8776905.

[4] P. Kim, *MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence*, 1st ed. USA: Apress, 2017.

[5] M. S. Elmahdy and A. A. Morsy, "Subvocal speech recognition via close-talk microphone and surface electromyogram using deep learning," in *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2017, pp. 165–168, doi: 10.15439/2017F153.

[6] Y. Liao and Y. Wang, "Some Experiences on Applying Deep Learning to Speech Signal and Natural Language Processing," in *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, 2018, pp. 83–94, doi: 10.1109/DISA.2018.8490638.

[7] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 International Conference on Engineering and Technology (ICET)*, 2017, pp. 1–6, doi: 10.1109/ICEngTechnol.2017.8308186.

[8] M. Z. Alom *et al.*, "The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches," Mar. 2018.

[9] R. Jagiasi, S. Ghosalkar, P. Kulal, and A. Bharambe, "CNN based speaker recognition in language and text-independent small scale system," in *2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 2019, pp. 176–179, doi: 10.1109/I-SMAC47947.2019.9032667.

[10] A. Antony and R. Gopikakumari, "Speaker identification based on combination of MFCC and UMRT based features," *Procedia Comput. Sci.*, vol. 143, pp. 250–257, 2018, doi: 10.1016/j.procs.2018.10.393.

[11] A. Tharwat, "Classification assessment methods," *Appl. Comput. Informatics*, vol. 17, no. 1, pp. 168–192, 2018, doi: 10.1016/j.aci.2018.08.003.

[12] J. Ma and L. Yang, "Robust supervised and semi-supervised twin extreme learning machines for pattern classification," *Signal Processing*, vol. 180, p. 107861, 2021, doi: 10.1016/j.sigpro.2020.107861.

[13] X. Zhang, D. Cheng, Y. Dai, and X. Xu, "Multimodal Biometric Authentication System for Smartphone Based on Face and Voice Using Matching Level Fusion," in *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, 2018, pp. 1468–1472, doi: 10.1109/CompComm.2018.8780935.

[14] A. Kamalu, A. Raji, and V. I. Nnebedum, "IDENTITY AUTHENTICATION USING VOICE BIOMETRICS TECHNIQUE U," 2015.

[15] R. Tanwar, K. Singh, and S. Malhotra, "An approach to ensure security using voice authentication system," *Int. J. Recent Technol. Eng.*, vol. 7, no. 5, pp. 161–165, 2019.

[16] S. Duraibi, F. Sheldon, and W. Alhamdani, "Voice Biometric Identity Authentication Model for IoT Devices," *Int. J. Secur. Priv. Trust Manag.*, vol. 9, pp. 1–10, May 2020, doi: 10.5121/ijsptm.2020.9201.

[17] N. Chauhan, T. Isshiki, and D. Li, "Speaker Recognition Using LPC, MFCC, ZCR Features with ANN and SVM Classifier for Large Input Database," in *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, 2019, pp. 130–133, doi: 10.1109/CCOMS.2019.8821751.

[18] S. A. Alim and N. K. A. Rashid, "Some Commonly Used Speech Feature Extraction Algorithms," in *from Natural to Artificial Intelligence - Algorithms and Applications*, IntechOpen, 2018.

Authors' information

Wahyu Ibrahim Received education at SMKN 56, Jakarta, in 2009, then continued with a bachelor's degree in engineering (S1), majoring in Electrical Engineering, Universitas Muhammadiyah Jakarta in 2012. I am a master of engineering student (S2) majoring in Electrical Engineering at Trisakti University. My research interests are in the field of electrical engineering and telecommunications engineering.

Henry Candra received a B.Eng. degree in 1995, a Master's degree in engineering in 2000 from Trisakti University, Jakarta, Indonesia, and a PhD degree in Engineering from the University of Technology, Sydney (UTS) in 2017. Currently, he is working as a Senior Lecturer and a researcher at the Electrical Engineering Department, Faculty of Industrial Technology, Universitas Trisakti in Indonesia. His research interests are in affective computing, brain-computer interfaces, biomedical signal processing, rehabilitation technology, and implementations of artificial intelligence and machine learning algorithms.

Haris Isyanto received a Bachelor of Engineering degree from the Department of Electrical Engineering Universitas Muhammadiyah Jakarta in 1996, a Master's Engineering degree from the Department of Electrical Engineering, Universitas Trisakti, Jakarta in 2000. He is a senior lecturer at the Department of Electrical Engineering, Universitas Muhammadiyah Jakarta Indonesia. His research interests are electrical engineering, wireless, 5G, IoT, electronics, and sensors.