International Journal of Gender, Science and Technology

..ISSN: 2040-0748

UGC Care Group I Journal

Vol-12 Issue-02 Nov 2023

Voice Recognition Authentication using CNN

Sabha Firdhos, Thombarapu Chandu, Gugulothu Vishal, Pavan Kumar Saroj,
Mr. A. Ramesh(Assistant professor),
Department of CSE,
MALLA REDDY INSTITUTE OF TECHNOLOGY AND SCIENCE, Telangana, Hyderabad.

Abstract –

In this research, we look at how the CNN method for deep learning may be used to assess the trustworthiness of speech recognition security. When it comes to learning, the CNN algorithm is superior since it is quicker, more accurate, and safer. CNN is also capable of resolving issues with user identification in datasets that are quite vast. A total of 10 distinct user voices, each having an iteration count of6,000,12,000, or 15,000, make up the measured voice input. In order to identify conversations and save crucial information, voice extraction features are also used. The next step in getting a trained model is training the data from the voice file iterations to record the user's voice. By comparing the real value with the predicted value in the CNN algorithm, these findings quantify performance (confusion matrix). Based on the findings, the optimal number of iterations for a sound file is 15000 iterations, which yields an accuracy of 96.87%. A lower number of iterations yields 96.30%, and a lower number of iterations yields 95.77%. According to CNN's performance statistics, a high level of accuracy is achieved after 15000 repetitions of audio files. Protecting one's identity and providing a high level of security are both made easier with voice recognition security.

I. Introduction

Cybercrimes like identity theft and data fraud pose a new danger to everyone's personal information. A human is crucial when it comes to identifying data and obtaining information. Passwords, magnetic cards, and PIN codes are the most prevalent techniques employed so far. Card abuse, damage, forgetfulness, loss, theft, hacking, and counterfeiting are some of the ways this falls short. Then they came up with a way to identify a person in order to lessen this issue [1]. For fast, accurate, private, and secure authentication and individual identification based on a person's biological traits, identity recognition technology is the way to go. Recognizance in a person is a physical trait that is distinct from all others and impossible to replicate. Hacking, reconstructing, and forging recognition techniques are far more challenging. in references [1], [2].

When it comes to knowing someone's identity, there are two things that make identification recognition work: their behavior and their physiological make-up. Some examples of physiological traits include vein patterns, eye shapes (including the iris and retina), fingerprints, hand shapes, facial features, and DNA. Gait, pulse rate, signature, keystroke dynamics, noise, and movement are all ways to observe behavioral traits linked to a distinct pattern [3]. The creation of a neural network to accurately perform tasks, such as object identification and voice recognition, is what deep learning refers to in artificial intelligence. Without pre-programmed rules or domain expertise, deep learning may represent a wide variety of data types, including video, text, pictures, and audio [4-6]. The needs of user identification challenges may be solved more safely, swiftly, and precisely with the aid of convolution neural networks (CNNs), which are a kind of data-driven approach [7]-[9]. When applied to user speech input, the convolution neural network algorithm yields accuracy metrics as a measure of performance.

When it comes to artificially applying the ear's functional principles, Mel-Frequency Cepstral Coefficients (MFCC) is the most effective and accurate way for detecting and extracting spoken sounds. By identifying talks, the MFCC can extract useful information while filtering out background noise [18]. To evaluate how well an algorithm compares the projected value to the actual value, or vice versa, one might use a confusion matrix. In order to generate accuracy parameters using the CNN algorithm, we quantify speech recognition voice input data in this work[10]-[12]. Voice recognition was selected for this study due to its many advantages: it can authenticate voices, it protects against fraud and other forms of fraud, it

has high security standards, and it maintains the privacy of personal identities. Additionally, voice recognition does not require special devices like retina scanners or fingerprint readers, which means that its implementation costs are lower. When it comes to person identification, the identifying recognition approach is both straightforward to use and accurate [1], [13]-[16].

II. Research Method

The voice recognition system used in this study collects user speech data from 10 different voice types and uses 6,000, 12,000, and 15,000 iterations to segment audio files. The collected voice data is subjected to feature extraction in order to preserve just the pertinent audio data and remove the irrelevant. Getting the voice ready for the training stage of the convolution neural network approach is the next unnecessary step. To do this, the user's voice must first be registered or labeled, which requires iterating over several audio recordings of speech. After the training stage is finished and a trained model is generated with user voice data recorded and saved in the database, the next step is authentication/verification, which entails evaluating the user's voice using two processes: speaker recognition and speech recognition. Get the chords to fit neatly. By comparing the recorded voice to the information held in the database about the trained convolution neural network (CNN) model, we were able to acquire both positive and negative results from the authentication process, which included speaker identification and speech recognition, respectively. In order to evaluate the performance of the convolution neural network method, we will next use a confusion matrix to compare the user's actual and anticipated voice values. For audio files with6,000,12,000, and 15,000 iterations, the accuracy value is determined by the results of the confusion matrix performance assessment, correspondingly. A schematic of the voice recognition system may be seen in Figure 1.

Stage two involves collecting data on user voice input, specifically 10 unique voices. An average sampling rate of sixteen thousand hertz (Hz) is used to record the user's voice. There are 6,000, 12,000, and 15,000 samples utilized. Training a model for a convolution neural network method to identify the user's speech follows feature extraction. Based on the anticipated outcomes of gathering user voice input, the system will be able to identify conversations and save relevant information. You may find the data that was gathered in Table I.

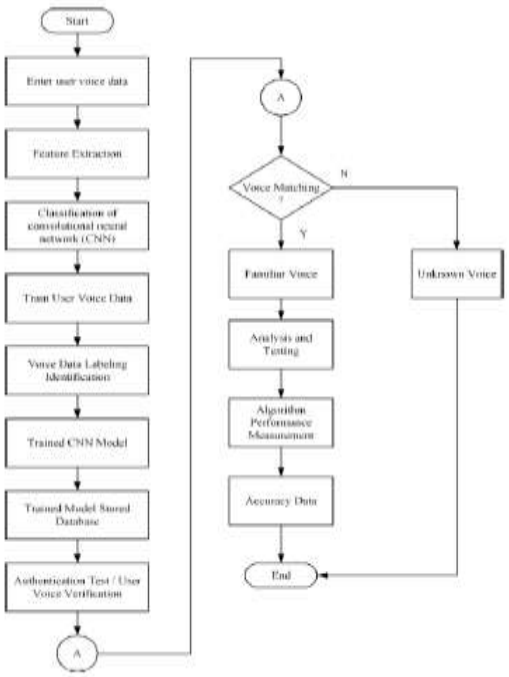


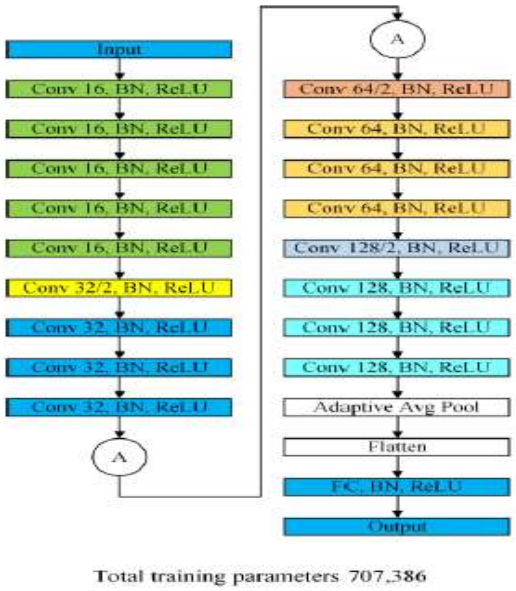
Figure 1: Voice recognition system flow diagram

DATA COLLECTION FROM VOICES (TABLE 1)

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**	

To train a speaker recognition system, one must first collect enough audio recordings. Then, a convolution neural network must be designed. Speech input data will be used as input by this algorithm. What follows is a network layer called ReLu that activates the activation function; batch normalization helps train on voice input data faster; a convolution layer with 16, 32, 64, or 128 convolution filters or kernels follows; and lastly, the adaptive average pool function determines the necessary kernel. Crucial for producing an outcome with the given dimensions. Connecting data to a fully linked layer follows flattening it into a one-dimensional array or line using the flatten function. After this layer applies batch normalization and ReLU to the incoming data, it determines if the data is legitimate or not. The CNN architecture is

accompanied with a total of 707,386 training parameters. Here is the CNN architecture shown in Figure 2.



Picture 2: The Architecture of a Convolutional Neural Network

Once one is acquainted with the architecture of convolution neural networks, they may proceed with training and authentication for voice recognition. It is necessary to handle the user's audio data (speakers) before training a convolution neural network algorithm to categorize voices. Verify if the CNN model's trained knowledge can recognize the input speaker as part of the authentication procedure.

As part of the training process, the convolution neural network algorithm is given user voice input. Training the data required varying total numbers of sound samples:6,000,12,00, and fifteen thousand. For a validation test ratio of 10%, training takes 40 epochs. The objective of the training process is to determine whether the speaker's voice is genuine or not. The final result of training models using user voice data is called a trained model. Table II displays the training data for the convolution neural network (CNN) model, which is stored in the voice recognition database.

Figure 2: Training with Condom

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

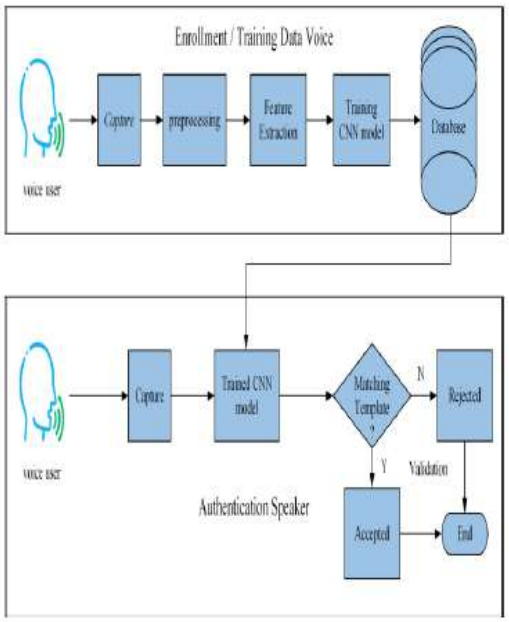


Figure 3: Speaker Recognition Block Diagram

The speaker recognition phase follows. Two processes are involved in processing speaker authentication data and registration/training data. After the user's speech is taken during registration, it goes through preprocessing to transform it into a digital signal, just like any other speaker's voice. Then, MFCC is used to extract the voice, which tries to identify the dialog so that it can remember certain details and ignore the rest of the noise in the signal. Ten distinct voices are used in the user voice data training procedure after feature extraction. After training the voice sample data, the results will be obtained in the form of a trained model. The amount of voice samples used for training may vary between 6,000, 12,000, and 15,000 files. The purpose of this training is to classify the user's voice as valid or invalid. The database will include the trained voices. The next step in speaker authentication is to identify the speaker by their voice. First, the user's speech is recorded. Then, the trained model's expertise is used to match their voice with a database. To allow for the validation process to approve or reject subsequent votes. Figure 3 shows the steps of speaker recognition.

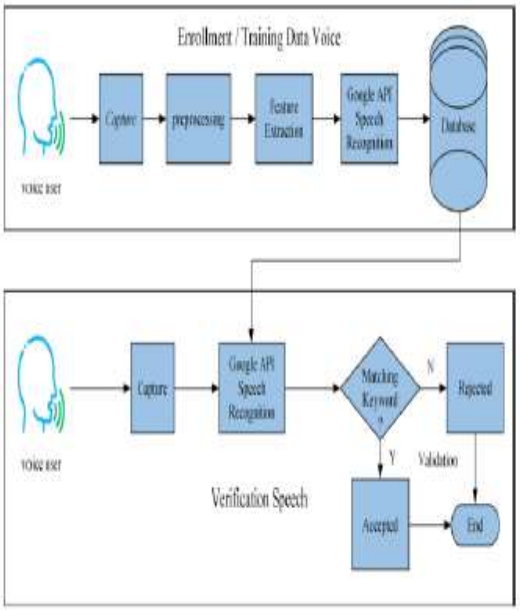
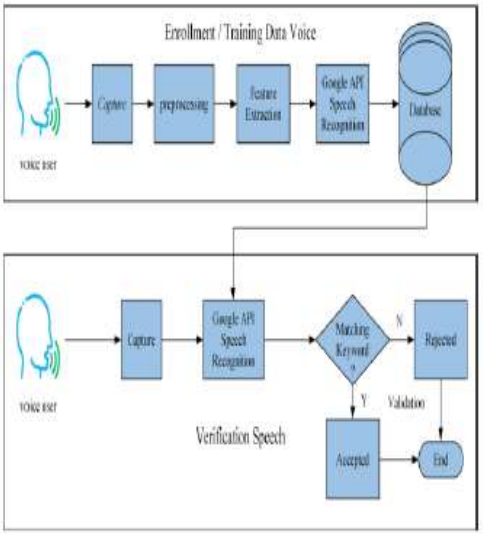


Figure 4: A Block Diagram of OCR

In addition, there are two steps involved in speech recognition: data registration/training and speech content verification. When registering, the first step is to record the user's voice as it is said by the speaker. Then, the sound is preprocessed to transform it into a digital signal. With the help of MFCC, which attempts to decipher the discussion in order to extract relevant information and exclude background noise, the voice is retrieved. After the features are extracted, the next step is to expose the security of the speech recognition system by registering terms using the Google API.



recognition of spoken language. Once a keyword is registered, it is saved in the database. Next, we'll use speech recognition to verify keywords in voice material. Sound is captured in order to carry out this procedure. Then, using what the user knows about the Google API's voice recognition, they speak the terms as registered keywords. The next step in the validation process involves matching the user-entered keywords with existing keywords in order to approve or reject them. Figure 4 shows the steps involved in voice recognition.

TABLE III: CNN Performance Evaluation Method

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

III. Result and Discussion

In order to find the best parameter comparison for each iteration, this research analyzed training and test data using 6,000, 12,000, and 15,000 sound files, respectively, to compare the accuracy value of the speaker identification algorithm on the convolution neural network algorithm model. The vocal, speech recognition, and response time tests check that the system pronounces phrases properly and measure how quickly it understands the **input voice.**

The examination of training models for convolution neural networks (CNNs) 3.1

Using6,000,12, and 15,000 audio recordings, this test will assess the iterative processing of the user's voice input data. Using metrics like accuracy and loss, the test will verify that the user's voice data training procedure is sound. The training technique consists of forty iterations. According to Table IV, which displays the data validation accuracy of the number of sound files, the CNN algorithm undergoes 40 cycles of training using user speech data.

Table IV: Validation of SFI Training Accuracy

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

The quantity of user voice recordings is the optimal data for the iterative data training method, as shown in Table IV. The best training data was obtained on epoch 38 with an accuracy validation value of 95.8852% in the iteration of 6,000 sound files; in the testing iteration of 12,000 sound files in epoch 32, the validation value of accuracy was 96.6235%; and in the testing iteration of 15,000 sound files in epoch 34, the validation value of accuracy was 97.4249%. Figure 5 shows a graph of training accuracy validation, which suggests that the user's speech data training process becomes more accurate as the quantity of sound files increases.

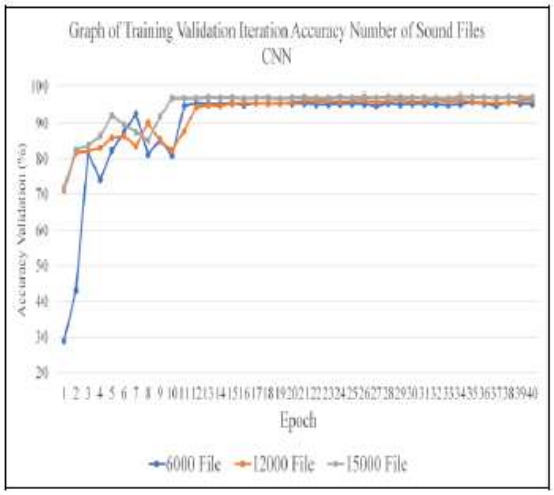


Figure 5: Graph Validating Training Accuracy

The next step is to evaluate the training procedure for the convolution neural network by measuring the loss validation. Using 40 epochs of training, the test is executed with iterations of 6,000, 12,000, and 15,000 sound files. In Table V, you can observe the following training test loss validation data.

VI. VALIDATION OF SOUND FILE ITERATION FOR TRAINING PURPOSES

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

Considering the data in table V. At epoch 21, we get the best data at 0.1398% for 6000 sound file iterations; at epoch 30, we get the best data at 0.1044%; and at epoch 34, we get the best data at 0.0891% for 15000 sound file iterations. At these instances, we get the loss validation values. Research on the loss validation parameter has shown that there is a positive correlation between the number of iterations and the loss validation value for 15,000 sound files. The training test on repetition of sound files is shown visually in Figure 6.

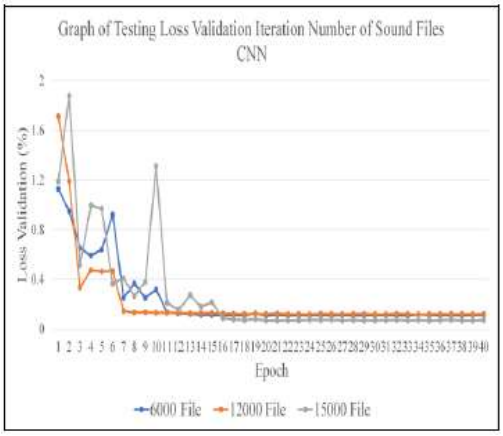


Figure 6, Validation Graph for Training Loss

By iterating through a large number of sound files, the best data is collected for accuracy validation and loss validation after the user's speech data is input into the convolution neural network (CNN) algorithm for training purposes. You can see the following comparison of training tests in Table VI, thus the data is created to identify the best outcomes in the training testing procedure.

The Top Data Testing, Training, and Validation Methods (Table VI)

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

As seen in table VI. A greater validation accuracy was seen when using more sound recordings to train the user's speech data. Validation accuracy in the 15000 sound file iteration test was 97.42%, in the 12000 sound file iteration it was 96.62%, and in the 6000 sound file it was 95.88%, as shown in Figure 7 of the accuracy validation graph.

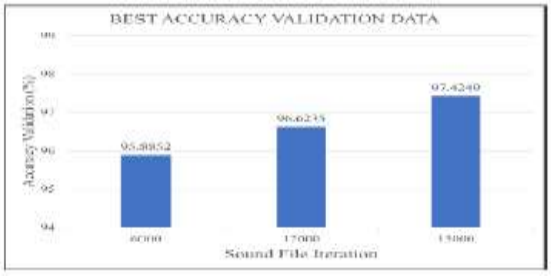


Figure 7: Graph for Validation of Best Accuracy

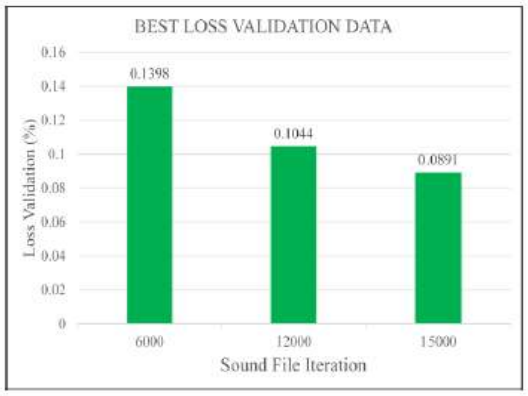


Figure 8: Optimal Graph for Loss Validation

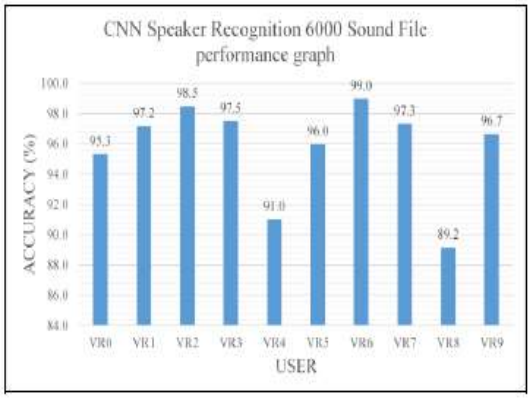
We go on to table VI. It was discovered that the lower the loss parameter, the more iterations of the sound file were used in the testing procedure. The loss percentage for sound files is 0.09% after 15000 iterations, 0.10% after 12000 iterations, and 0.14% after 6000 iterations. Figure 8. III displays the optimal loss validation graph.2. Evaluation of CNN's ability to recognize speakers

This research evaluated the efficacy of an algorithm based on convolution neural networks. Voice input data speaking (speaker) is the format of the measured data. Iterating between 6,000, 12,000, and 15,000 audio recordings allows users to compute the projected value and actual value. Ten different user voices' input data is measured by the test. Ten percent of the total file iterations are used as data to test algorithm performance. Later votes will be used to determine CNN's performance accuracy in iterations, after being aware of the predicted value parameters and the actual values in the form of TP, FP, FN, and TN to demonstrate similarities to other users' voices. Several files. In table VII, you can see the results of the following algorithms.

Performance Test with 6,000 Sound Files (Table VII)

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

Table VII shows the results of calculating the actual and predicted values of each user's voice data. Very high false-positive (FP) and false-negative (FN) results indicate that users of VR0, VR4, VR5, and VR8 are comparable. In this situation, it may lead to less precise results. In addition, VR6 has the best accuracy value (99%), as it has the lowest FP and FN values. Figure 9 shows a graph of the performance measurement for speaker recognition.



Graph showing the performance of CNN speaker recognition on 6,000 audio files (Fig. 9).

IV. Conclusion

1Training parameters and performance are both improved with an increased number of iterations for the audio file.

The results of the training test showed that at iteration 15000, the validation accuracy score was 97.42% and the validation loss value was 0.891%.

3. Applying a confusion matrix to evaluate the CNN algorithm's performance across 15,000 audio recordings yields the following results:

The highest degree of precision that can be attained is 96.87%. With 12,000 audio files iterated, the accuracy is 96.30%, but with 6,000 audio files iterated, it drops to 95.77%.

Twenty trials later, the voice recognition system had a 95% success rate with a 5% failure rate.

5. The results of the test for voice recognition indicate that it typically takes 3.85 seconds to understand the user's speech.

6. The voice recognition system is very applicable due to its two authentication measures: speech recognition and voice content verification.

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.