[1]Ashraf Tahseen Ali [1], Jamshid Bagherzadeh Mohasefi [2]

Fraud Recognition Using Voice Authentication Through Deep Learning Applying 1D-CNN Algorithm.

Journal of Electrical Systems

Abstract- The identification of individuals has become one of the most significant issues in the protection of any system, this is because of the spread of the Internet, computers, and electronic accounts. The recognition of users through biometric features is one of the fundamental principles, and voice is one of the most effective ways to recognize... Compared to other traditional biometric methods, they offer a similar level of security and are dependable and trustworthy. The objective of this research is to explore deep learning using the proposed one-dimensional convolutional neural network (1D-CNN) on two different voice datasets: natural voice recordings and unconstrained voice recordings. The dataset that has the greatest success in recognizing and categorizing speakers is the voice recording that is taken in a natural environment. To enhance the quality of the acoustics' in real-world environments, we pre-process the audio using noise reduction and enhancements based on Mel-Frequency Cepstral Coefficients (MFCC) for each sound, as well as their variance and acceleration. Our first attempt at training and testing the 1D-CNN, the results showed that it had a 100% success rate.

***Keywords*:** Biometrics; Voice recognition; Machine learning; Deep learning:1D-CNN.

1. INTRODUCTION

utilizing two-step verification in one procedure. The process of recognizing voice is of paramount importance to safe and unique biometric authentication, although other methods of identification are also employed. Personal voice recognition and telephone recognition are both dependent variables that can be altered by speech. Voice recognition systems are inexpensive and simple to use. In today's smart world, voice recognition is critical in several ways. Voice recognition has a wide range of applications, including voice-started banking, Internet of Things [1]. The voice is one of the ways to identify someone. The speaker recognition system assesses a person's condition based on their voice, including gender, accent, speech, emotion, and health status. Along with technological improvements, it has practical applications, such as voice-based security systems [2]. Preprocessing is the initial stage of other levels in speaker recognition that separates voiced or other signals and generates feature vectors. The important and common steps include noise removal, pre-emphasis, voice activity discovery, windowing, and framing. It should be noted that the method of processing speaker recognition varies depending on the type of data set and its format (wave or mp3), with both being the most common. Given the amount of noise in each audio clip, preprocessing is required to remove noise and useless data. Researchers used a variety of methods for feature extraction, including MFCC which is the most common and widely used, to achieve their goals by applying appropriate parameters. Employing a deep learning classifier that produces effective outcomes. In a nutshell, these are the most crucial elements in the process of ID for the speakers [3].

[1] Email:ashraf88ashraf888@gmail.com 1
Email:j.bagherzadeh@urmia.ac.ir 2
[1]Department of Computer Engineering, Urmia University, Urmia, Iran,

2. LITERATURE SURVEY

Speaker Recognition (SR) is a biometric approach, similar to other biometrics like finger veins, palm, retina, iris, and facial recognition, that uses an automated system to identify individuals based on their voice signal. The primary characteristic that sets (SR) apart from other biometrics is that contrary to other approaches that typically rely on picture information, (SR) can be found to be the only technology that prepares spoken data [3]. Some of the earlier research-related material will be covered in this section:

Naveen, et al. [4], suggested Every audio feature was given its own Dense Neural Network, which was then concatenated using a concatenation layer to yield the best performance output when compared to LSTM. Most of the time, the speaker was correctly predicted by the Dense Neural Network with an accuracy of 86%. Noor, et al. [5], proposed a deep learning-based model (DNN) with an architecture to identify the individual by taking advantage of the distinct personal traits present in each person's voice. To add more samples to the existing dataset, an augmentation approach is applied. After analyzing the temporal data provided in an input audio file, feature maps representing the prominent temporal features (time-domain features) are derived from the data. The choice is taken after keeping track of these vocal characteristics over time. For the identification of 40 participants, the accuracy over the VoxCeleb1 dataset is around 99.81%. Jolie, et al. [6], suggested using convolutional neural networks (CNNs) to identify individuals from audio databases. The 60 participants in the self-created database were recorded using voice recognition software. Spectrophotograms from audio databases are utilised to do speaker recognition on two networks: Inception v3 and MobileNet v1. This assignment makes use of a dataset of 12,000 spectrogram images, of which 9600 are used for training, 1200 for network validation, and 1200 for testing. The MobileNet v1 network is surpassed by the Inception v3 network, which offers an accuracy of 85.5%. Giovanni et al. [7], investigated and contrasted the two approaches using the voice dataset, which included 8869 audio recordings of 58 speakers. Spectrophotom and Cepstral-temporal (MFCC) graph picture inputs are used to assess a bespoke CNN against many pre-trained nets. An AML method that utilises a Naïve Bayes model for feature extraction, selection, and multi-class classification is also taken into consideration. The most accurate findings, 90.15% on grayscale spectrograms, are obtained by using a custom, less deep CNN that was trained on grayscale spectrogram pictures. Sung1, et al. [8], To increase speaker recognition accuracy, they developed a convolutional neural network (CNN) in conjunction with an acoustic model of Connectionist temporal classification. They also proposed the construction and training process of the acoustic model Connectionist temporal classification (CTC) algorithm. The outstanding performance of the CTCCNN_5 + BN + Residual model structure was further confirmed by comparison verification using the speaker's speech rate, modelling units, and acoustic feature parameter choices. With decreased noise pollution, the accuracy of the CTC-CNN baseline acoustic model reached 97.33% after 54 training epochs using the THCHS-30 and ST-CMDS speech data sets as training sets. Xinhua, et al. [9], identified the speakers for the first time, the dung beetle optimised convolution neural network (DBO-CNN) was presented, which helps in determining appropriate hyperparameters for training. The 50-person dataset was tested, and the results showed that this method greatly increased the model's accuracy. In contrast to conventional CNN and CNN enhanced by additional sophisticated algorithms, DBO-CNN has an average accuracy of 97.93%.

3. THE PROPOSED METHOD

The suggested biometrics-based system would recognise speech using (CNN) algorithms. This section includes the following: the description of the database; preprocessing; feature extraction; hold-out cross-validation stages; and classification stages. The proposed method design is shown in Figure(1).

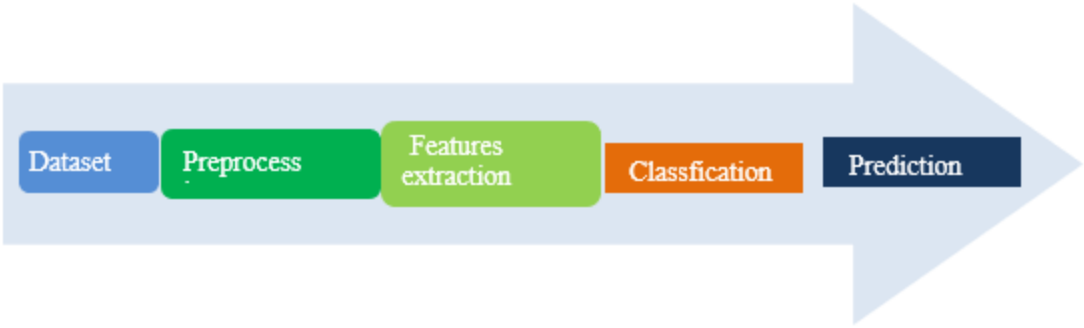

Figure 1. The proposed voice recognition system planning

3.1. Database Review

Right now, the dual most widely used audio formats are WAV and MP3. The majority of researchers use WAV files because they contain all of the frequencies that are perceptible to the human ear. Conversely, MP3 files are compressed and do not have all the information found in a WAV file containing comparable audio. Moreover, it is essential to extract functions from these WAV files. The machine learning techniques that will be employed to categorise the data are built upon this step. WAV files are therefore frequently utilised in audio investigations. To guarantee that the recovered coefficients in an audio sample accurately reflect the underlying computations, sampling rate consistency is essential. The details of this data are displayed in Table.1 The files for both datasets with their details were obtained from Kaggle.

Table 1. Datasets specifications

Input file Data	Name of Dataset	File format	Noise degree	No. of sample
Voice Recognition	Prominent leadersspeeches	Wave	high	7500
	Speaker Recognition	Wave	low	2226

3.2. Pre-processing

The primary benefit of the preprocessing step is that it arranges the data, which facilitates the recognition process. "Preprocessing" refers to any audio-related activities, including "noise removal" utilising the Hamming Window (HW). It chooses a sufficiently representative slice to analyse lengthy sound signals. This procedure is used to eliminate noise from a signal that is contaminating the current frequency spectrum.

after that, utilising Speech signal processing, which is based on the time domain input/output relationship shown in the following equation, requires the pre-emphasis filter. Smoothing the unique form of the speech signal frequency is the goal of employing this filter.

The next step applies To express a function with finite time, utilise the Fourier series. The Fourier transform is utilised to convert a time series of signals with restricted time-domain into a frequency spectrum. The time domain to frequency domain conversion of each frame is accomplished by this procedure.

3.3. Feature Extraction

The process of computing a set of feature vectors that yields a condensed representation of a certain speech signal is known as feature extraction. Utilise Mel-Frequency Cepstral Coefficients (MFCC), a technique that detects frequencies higher than 1 kHz by utilising the activity of human hearing. The frequency changes that are perceptible to the human ear are the main focus of the MFCC system. The suggested voice system is more sophisticated because 12 Cepstral Coefficients were chosen, as opposed to Vector Quantisation in Practice which, in the digital representation of signals for computer processing, is an inevitable stage. In this case, it was utilised to change the binary matrix produced by MFCC into a one-row matrix, which it then merged with the output matrices from the other tools.

3.4. Proposed System Classifiers Using (1D-CNN)

(1D-CNN) is a neural network that operates on one-dimensional data, such as time series or sequence data. In this study, 1D CNNs are used to extract representative properties from corona and non-corona faults in both the time and frequency domains. This is achieved through 1D convolution operations using filters. The Fourier series can be used to express a function with a finite time. A time series of bounded time-domain signals is converted into a frequency spectrum using the Fourier transform. This process is used to convert each frame from the time domain to the frequency domain.The convolutional layers, filters, MaxPooling1D layer, FC layer, and classification layer with a Rectified Linear Unit (RELU) as the activation function are all part of the CNN structure. To prevent overfitting, batch size and dropout are employed. There exist statements for these equations. (RELU) is a CNN activation function that has the potential to generate nonlinearity. Deep CNNs have been applied to video analysis and recognition applications, while their original application was in image classification. On the other hand, 1D sequence data for prediction and classification is still relatively new. 1D CNNs are a good option since the corona and non-corona categories may be thought of as a sequential modelling effort. As demonstrated by Algorithm 1, compressed 1D-CNNs are favoured for real-time applications because of their minimal processing requirements.

Algorithm 1. 1D Convolutional Neural Networks (CNN)

Input: The split dataset
Output: accuracy of CNN
Begin
1. consists of the first convolutional layer in our CNN architecture, the convolutional layer with the Leaky ReLU activation function. Three is the size of the kernel, one is the padding (p) around the entire image, one is the stride (s), and sixteen filters are used. MFCC and three stages of feature extraction make up the input shape.
2. layer of maximum pooling. The previous layer provides the size 16 input to this layer. Padding is 1, stride is 2, and pooling size is 2*1.
The Leaky ReLU activation function in the second convolutional layer. The input of size 32 is passed from the previous layer to this one. Padding is 1, stride is 1, and filter size is 32..
Max Pooling Layer. The input of size 32 is sent to this layer from the layer above. Padding is one, stride is two, and pooling size is two by one.
the Leaky ReLU activation function in the third convolutional layer. The input of size 64 is passed from the previous layer to this one. The padding is 1, the stride is 1, and the filter size is 64.
layer of maximum pooling. The input of size 64 is passed from the previous layer to this one. Padding is 1, stride is 2, and pooling size is 2*1. Following this max pooling procedure
. the Leaky ReLU activation function in the fourth convolutional layer. The input of size 128 is sent to this layer from the layer above. The padding is 1, the stride is 1, and the filter size is 128.

. Max Pooling Layer. The previous layer provides the 128-size input to this layer. The dimensions of the pooling are 2*1, the padding is 1, and the stride is 2.

9. The fifth convolutional layer with Leaky ReLU activation function.This layer gets the input of size 256 from the previous layer. The filter size is 256; padding is 1, the stride is 1, and the number of filters is 256.

10. Padding is one, stride is two, and the maximum pooling size is two by one. Following this max pooling process, 512-size feature maps are obtained. Every feature map undergoes separate max pooling processes.

1. The activation function of the sixth convolutional layer is a leaky ReLU. The input of size 512 is sent to this layer from the layer above. Padding is 1, stride is 1, and filter size is 512.

12. layer of maximum pooling. The input of size 512 is sent to this layer from the layer above. Padding is 1, stride is 2, and pooling size is 2*1. following this max pooling procedure.

3. The activation function of the seventh convolutional layer is Leaky ReLU. The input of size 1024 is sent to this layer from the one above. The padding is 1, the stride is 1, and the filter size is 1024.

14. The eighth convolutional layer has an activation function of linearity. The previous layer provides the input of size 1024 to this layer.

5. Flattening converts the data into a one-dimensional array so that it may be entered into the following layer. The output of the convolutional layers is flattened to produce a single lengthy feature vector. It is associated with the last classification model, called a fully-connected layer.

6. The network's final layer is called dense. It is an output layer that is entirely connected. Their input size ***n***, where n numbers represent a class score, for example, among the dataset's ***n*** categories. We employ the Softmax activation function for the final results.

End

4. THE PROPOSED SYSTEM IMPLEMENTATION

The two datasets are trained and testing using holdout cross-validation in the proposed system. testedThe preprocessing of the dataset will involve the application of a hamming window. Subsequently, the features will be extracted and modelled through the utilisation of MFCC and VQ. The values of these features will be amalgamated with the extracted features in order to be ready for the classification algorithms. Throughout the training phase, the mixed features are stored as reference models. After that, these models are contrasted with the input voice signals. The proposed system architecture is shown in Algorithm 2 below.

Algorithm 2. The proposed voice recognition system

Input :- Voice datasets

Output: - Classifier performance

Begin

1. Load two voice dataset
2. Eliminating noise by means of the (HW)

Pre-emphasis used to smooth down the speech signal frequency's spectral form

4. Signals adapting into a frequency spectrum applying Fast Fourier Transform (FFT)

Human hearing activity is used to calculate Mel-Frequency Cepstral Coefficients (MFCC) to identify frequencies.

Digital illustration of signals applying Vector Quantization (VQ)

7. Split the dataset arbitrary and divide it into (training and testing) groups using Hold out Validation

8. Classify instances based on 1D CNN **Call algorithm 1**

9. Classifiers Evaluation

End

5. PROPOSED SYSTEM EVALUATION

For evaluating a model's performance, certain parameters are used to determine its behaviour. The results are influenced by the size of the training data, the quality of the audio files, and, most importantly, the type of machine-learning algorithm used. The following criteria are used to assess the models' efficacy [10]:

Accuracy: Percentage of cases correctly categorized from all given examples. It is calculated as:

$$\text{Accuracy} = \frac{ta+tb}{tb+tb+fa+fb} \ [10] \ . \ (6)$$

Precision: The percentage of true x-class instances for all those listed as class x. It is calculated as:

$$\text{Precision} = \frac{ta}{ta+fa} \ [10] \ (7)$$

Recall: The percentage of cases listed as class x among all examples of class x. It is calculated as:

$$\text{Recall} = \frac{ta}{ta+fb} \ [10] \qquad (8)$$

F- measure: is the harmonic mean of precision and recall. It is calculated as:

$$F_1 = 2 * \frac{precision*recall}{precision+recall} \ [10] \qquad (9)$$

Where

ta = true positives: the amount of cases predicted positive that are positive

fa = false positives: the amount of cases predicted positive that are negative

tb = true negatives: the amount of cases predicted negative that are negative

fb = false negatives: the amount of cases predicted negative that are positive

Error rate: An error is merely a misclassification, as demonstrated by the following Eq. (10), when the classifier presents a case and improperly classifies it:

$$\text{Errorate} = 1 - \text{accuracy} \ [10] \ (10)$$

Specificity: quantifies a test's capacity to be negative in the absence of the condition. It is also referred to as the null hypothesis, Type I error, α error, false-positive rate, accuracy, and error of commission.

$$\text{Specificity} = \frac{tb}{tb+fa} \ 100\% \ [10] \ (11)$$

6. EXPERIENTIAL RESULTS

In this experiment, we use a (1D-CNN) classifier to test our datasets. Table (2) will show the outcomes of using the voice (dataset1, dataset2) as input. Each of the five individuals in the first dataset has 1500 samples, while each of the 50 individuals in the second dataset has between 40 and 50 samples. Table 2 displays the f-measure, specificity, accuracy, precision, and recall.

Table 2. Results of 1D CNN Classifier

	1D CNN	
	Dataset1	Dataset2
Total instances	7500	2226
Total correct	7500	2226
Total incorrect	100	100
Total Accuracy	100	100
Total Precision	100	100
Total Recall	100	100
Total F- measure	100	100
Total Error rate	0	0
Total Specificity	100	100

7. CONCLUSIONS

human voices. We believed that the analysis would be enough if we only used the MFCC coefficients. The proposed deep learning method and two datasets were employed. According to our research, accuracy was enhanced when a deep learning classifier was used in the classification process; 1D CNN classifiers were able to achieve 100% accuracy. It was better than the earlier work's accuracy scores.

REFERENCES

[1] Ali, A. T., Abdullah, H. S., & Fadhil, M. N. (2021). Voice recognition system using machine learning techniques. *Materials Today: Proceedings*, 1-7.

[2] Ali, A. T., Abdullah, H., & Fadhil, M. N. (2021). Speaker Recognition System Based on Mel Frequency Cepstral Coefficient and Four Features. *IRAQI JOURNAL OF COMPUTERS, COMMUNICATIONS, CONTROL AND SYSTEMS ENGINEERING*, *21*(4), 82-89.

[3] Ali, A. T., Abdullah, H., & Fadhil, M. N. (1970). Impostor recognition based voice authentication by applying three machine learning algorithms. *Iraqi Journal of Computers, Communications, Control and Systems Engineering*, *21*(3), 112-124.

[4] Naveen, R., Reddy, J., & Tanguturu, R. (2022, August). Speaker Identification and Verification using Deep Learning. In *2022 International Conference on Signal and Information Processing (IConSIP)* (pp. 1-6). IEEE.

[5] Al-Shakarchy, N. D., Obayes, H. K., & Abdullah, Z. N. (2023). Person identification based on voice biometric using deep neural network. *International Journal of Information Technology*, *15*(2), 789-795.

[6] Gomes, J., Fernandes, H., Abraham, S., & Chavan, S. (2021, January). Person identification based on voice recognition. In *2021 4th Biennial International Conference on Nascent Technologies in Engineering (ICNTE)* (pp. 1-5). IEEE.

[7] Costantini, G., Cesarini, V., & Brenna, E. (2023). High-level CNN and machine learning methods for speaker recognition. *Sensors*, *23*(7), 3461.

[8] Sung, W. T., Kang, H. W., & Hsiao, S. J. (2023). Speech Recognition via CTC-CNN Model. *Computers, Materials & Continua*, *76*(3).

[9] Guo, X., Qin, X., Zhang, Q., Zhang, Y., Wang, P., & Fan, Z. (2023). Speaker recognition based on dung beetle optimized CNN. *Applied Sciences*, *13*(17), 9787.

[10] Ali, A. T., Abdullah, H., & Fadhil, M. N. (2021). Speaker Recognition System Based on Mel Frequency Cepstral Coefficient and Four Features. IRAQI JOURNAL OF COMPUTERS, COMMUNICATIONS, CONTROL AND SYSTEMS ENGINEERING, 21(4), 82-89.