IETE JOURNAL OF RESEARCH
2024, VOL. 70, NO. 5, 4385–4398
https://doi.org/10.1080/03772063.2023.2252392

An Efficient Voice Authentication Approach Using Improved Deep Convolutional Neural Networks with LSTM Networks

N. Kaladharan* and R. Arunkumar

Department of Computer Science and Engineering, FEAT, Annamalai University, India

ABSTRACT

Voice authentication and verification play an essential role in forensic science and security systems in today's technological environment. A challenging issue is making precise voice authentication and verification. This document aims to identify scammers using machine learning techniques to find the best speaker identification and classification combination. Traditionally, researchers have approached voice authentication and speech comparison using separate classification and comparison models. However, the accuracy of training with large amounts of data has made deep learning more popular. In this study, we predict speech recognition in speech datasets using an improved deep convolutional neural network (IDCNN) and long short-term memory (LSTM) network structure. The proposed approach entails three main steps: pretreatment, feature extraction, and classification. We could predict voice authentication and verification in this study with an accuracy of 99.3%. Aside from comparing classification performance, performance metrics for accuracy, sensitivity, and specificity were also computed. The newly introduced methods offer superior error rates, precision, F-score accuracy, and recall compared to conventional algorithms. This research is regarded as a pioneer in the field as a quick and effective way to enhance continuous speech recognition activity.

KEYWORDS

Improved DCNN (IDCNN); Long short term memory (LSTM); VGG-16 and CNN

1. INTRODUCTION

Sensitive data needs to be protected, and people must be verified before accessing it. Personal information is typically maintained in a distributed way, especially with the rise in the number of offers for various web portals, telephone services, and mobile devices. The most common form of authentication is through IDs or passwords. Protecting data against ever-increasing processing power and sophisticated password surveillance software must be longer and more complex. For identification, they use biometric traits like fingerprints, the iris of the eye, the face, the voice, or other biometric characteristics or behavioral patterns. Although not as distinctive in each person, these characteristics are unique to one person. To authenticate the user, biometric speaker verification recognizes voice functions.

The primary and most natural mode of interpersonal communication is speech. People typically have no issues adopting language as a biometric due to this and the fact that it is the primary form of personal identity (PI). The benefits of employing speech as a biometric include its simplicity, user-friendliness, hands-free and eyes-free operation, ease of application to allow remote recognition (on the telephone, Internet, *etc.*), and implementation cost, which is frequently low (frequently requiring only software) [1,2].

Training, registration, and evaluation are the general stages of voice authentication. To learn the individual characteristics of the speaker from the vocal cues, the system is programed using the available data during the training phase. The format system receives the speaker's expressions during the enrollment phase to create a model of the speaker. Then during the evaluation phase, a pronunciation model of the test speaker is created and contrasted with the current models to determine how similar it is to speakers who have already registered [3]. Access to dial-up agents, telephone banking, dataset access services, information services, voicemail, security checks for sensitive information, and access to remote devices will be monitored [4].

Depending on the kind of information utilized for registration and identification, there are two different kinds of speaker verification systems: text-dependent and text-independent modes of operation. The speaker's spoken text remains the same for recording and evaluation in text-dependent recognition systems. Text-independent systems allow for the recording and evaluation of a

*Present address: Lecturer, Department of Computer Engineering, Government Polytechnic College, Theni, Tamil Nadu.

variety of spoken texts; these systems are generally used to locate and authenticate speakers [5].

1.1 Problem Statement

Voice authentication is a computationally demanding task. A fully connected network is required for this. However, creating a fully connected speech network is a computationally challenging task. Speech is typically composed of noise and external factors that influence speech quality. Poor voice quality can cause output issues. Training a system for correct voice authentication and testing it for the accuracy and consistency of recognized characters can make voice authentication a computationally intensive task [6].

Traditional voice authentication methods and speech comparison systems, such as HMM (Hidden Markov Model), GMM (Gaussian Mixture Model), and VQ (Vector Quantization), make use of unique features of the speaker's speech characteristics. However, these methods are unsuitable for text-independent speech comparison systems and only for small-scale speech recordings. Furthermore, these techniques are computationally expensive when dealing with large distributions and data sets.

This disadvantage is that SVM in voice authentication systems performs poorly when the number of functions exceeds the number of samples. Standard CNN is superior to SVM regarding image input because it recognizes critical features without human intervention. However, standard CNNs can only succeed when the dataset is more extensive. CNN is prone to overfitting due to the model structure's complexity.

The recurrent neural network (RNN) has significant advantages in improving the accuracy of voice authentication. However, some RNN-based intelligent voice authentication applications need to be improved in protecting the privacy of voice data. In contrast, other privacy-preserving applications are highly time-consuming, particularly in model training and voice authentication.

Long Short-Term Memory (LSTM) neural networks are well-known for high accuracy and practicality. Although LSTM significantly improves the commercial viability of voice data and the accuracy of voice authentication while protecting voice data privacy, it is still scalable in terms of semantic dependencies and time consumption.

MLPs can handle highly complex tasks, but their training is time-consuming and computationally intensive.

1.2 Major Contribution

- This work first denoises the input speech signal using a recursive least-squares adaptive filtering method and a threshold-based denoising method.
- We extract low-level descriptor features from audio files, and then use a VGG-16 blocks-based baseline model to understand better the high-level hidden local features extracted from those extracted features during training, and finally classify the speech signal.
- To implement the voice authentication system using hybrid IDCNN-LSTM.
- The proposed IDCNN-LSTM models' performance is compared to previous existing models.
- The experimental results show that the IDCNN-LSTM hybrid algorithm is computationally efficient and competitive with existing reference systems.

This paper covers the long short-term memory (LSTM) application in voice authentication. For voice authentication, LSTMs are utilized one at a time, starting with feature extraction and ending with the entire system. Long short-term memory (LSTM) recurrent neural networks (RNNs) is spatially and temporally localized, and it is closely similar to the biological pattern of memory in the prefrontal region [7]. They have not only beaten the prior artificial RNNs in numerous artificially manufactured sequential processing tasks but are also more biologically believable. We used LSTM to solve more practical issues, such as speech recognition of numbers.

Because the spatial and temporal properties of speech signals are critical for high authentication rates, combining two separate networks were considered. To properly utilize these features and improve continuous voice authentication tasks, hybrid IDCNN-LSTM architecture is proposed in this research. The voice authentication task demonstrates the benefit of the IDCNN-LSTM architecture. This research aims to create a deep hybrid structure for acoustic modeling that addresses the leakage gradient problem while also taking advantage of prior knowledge. After examining the input feature set, various nonlinearities, and IDCNN architectures, a hybrid structure of IDCNN and LSTM is developed as a solution. IDCNN controls the translation variance and LSTM, solving the vanishing gradient problem. IDCNNs extract data from speech frames, which LSTM layers process in both directions. The improved model performance is due to using LSTM modeling capacity to directly process speech signals and exploit past information to estimate speech parameters reliably. The findings of this research demonstrate the promising potential of the LSTM model as a well-known technique in continuous voice authentication that can be implemented as a

hardware model for controlling stand-alone electronic devices.

The remainder of the article is organized as follows. The recent literature survey is discussed in section 2. Section 3 presents the proposed Improved DCNN with the LSTM algorithm for voice authentication. The experimental results and analysis is described in section 4. Finally, Section 5 presents the conclusions.

2. LITERATURE SURVEY

Navya Saxena *et al.* [8] proposed intelligent home security solutions using artificial neural networks for speaker and facial recognition. The user can watch his home using a computer, tablet, or mobile phone with the assistance of the suggested application. Face recognition is accomplished using this technique by recording a live feed of the person at the door, followed by a live feed analysis in which the detected face is verified using the owner's information stored in the database. The suggested model's total accuracy is 82.71%, with facial authentication accuracy of 87.5% and speech authentication accuracy of 84.62%. The proposed models outperform state-of-the-art models, which need more extensive training datasets for small datasets.

Hossein Salehghaffari *et al.* [9] suggested Convolutional Neural Networks for speaker verification. This study modeled convolutional neural network architecture for speaker verification to collect and discard data from speakers and non-speakers. During the training phase, the network is programed to differentiate between several speaker identities to build the background model. This creates a system that simultaneously gathers speaker-related data and builds stability for speaker-specific fluctuations. Traditional verification techniques that make speaker models directly from the background model have been proven ineffective compared to the proposed method.

Md. Rayhan Ahmed *et al.* [10] four novel architectures Model-A (1D CNNs-FCN), Model-B (1D CNNs-LSTM-FCNs), Model-C (1D CNNs-GRU-FCNs), and Ensemble Model-D, which integrates Model-A, Model-B, and Model-C using a weighted average methodology. We employ five standard reference data sets to assess the performance of these models: TESS, EMO-DB, RAVDESS, SAVEE, and CREMA-D. All four models were discussed to perform remarkably well at the SER task of identifying emotions from the raw sound of speech. The source code for the four suggested structures is accessible from the following GitHub repository to enable future expansion and repeatability of the findings.

Salahaldeen Duraibi *et al.* [11] examine the utility and practicality of authentication of IoT users using voice recognition techniques. It explores the methods and techniques used to create voice recognition systems and discuss whether they are appropriate for IoT surroundings. A voice recognition system is suggested to verify users of the IoT ecosystem.

Jeeweon Jung *et al.* [12] developed a D-vector-based speaker verification system using Raw Waveform CNN. To replace the extraction and preprocess acoustic characteristics with the raw CNN audio, d- vectors produced from the sample-level raw audio-CNN were investigated in this study. Traditional techniques for extracting acoustic features, including Mel-frequency spectral coefficient and Mel-filter bank characteristics, are no longer required because raw audio-CNN accepts raw waveform signals as input. According to the short-sentence experiment results in part 1 of the RSR2015 dataset, EER was decreased using raw audio-CNN from 8.34% to 7.61%.

Bancroft *et al.* [13] proposed to investigate how speaker verification and emotion recognition interact. The infrastructure to study these complex problems is established in this mission. It also includes exploratory studies to analyze the speaker reliability of existing dynamic detection systems for specific emotional behaviors automatically collected from target speakers. The experimental comments yielded promising results, with most recovered comments directed to the target speaker and expressing desirable emotions. The results also performed well for the dynamic recognition task, showing a 45.8% transfer to the target region of the stimulus valence space and a 77.4% transfer to the target quadrant.

T. Muruganantham *et al.* [14] created a Biometric Of Speaker Authentication Using CNN. The individual voice is collected as a data source, and the MFCC (Mel Frequency of Cepstral Coefficients) calculation is used to determine the distinctive coefficients in a particular sample. CNN is then used to prepare for the voice exams. Convolutional neural networks are fundamental neural architectures that use convolution at any layer rather than general lattice augmentation. The testing technique starts after the preparation phase is finished. It identifies a suitable candidate that stands out and performs properly during the testing procedure if the prepared voice tests are given as information.

Heinz Hertlein *et al.* [15] proposed Effectiveness in the Realization of Speaker Authentication. Experimental investigations are carried out based on the use of digital expressions derived from the XM2VTS database. The document thoroughly explains each unique strategy taken into consideration and presents the empirical findings in light of the various circumstances.

The experimental research discussed in this paper focused on the first aspect, mainly how the paradigm used to create the training and test materials affects the scenario's correctness. The findings unmistakably demonstrate a decline in verification accuracy due to a reduction in the training material's textual content (*i.e.* in some cases, a decrease in training data as low as two-digit sentences).

Feng Cheng *et al.* [16] experiments a Visual speaker authentication with random prompt texts by a dual-task CNN framework. The three functional components of the new DCNN presented in this paper are the lip feature, recognition, and content network. The system adopted several 3D residual units that accurately depict lip biometrics' static and dynamic properties. The experimental results demonstrate that, when compared to several state-of-the-art methodologies, the suggested approach performs better in fixed passwords.

Mohit Dua *et al.* [17] proposed LSTM and CNN-based ensemble approaches for spoof detection tasks in speaker identification systems. The work provided in this document aims to address this issue by applying deep learning (DL) techniques and various neural networks. The first model combines long short-term memory layers (LSTM) and time-distributed compact layers. The other two DNNs use spatial convolution and temporal convolution as their foundations. In addition, an ensemble model using these three DNNs was examined. The suggested set performs better with CQCC on the interface when tested with the 2015 ASV spoof dataset. Additionally, on evaluation sets, the system trained using the ASV spoof 2019 dataset outperforms the baseline system.

Xuesong Gao et. [18] suggested a home automation device verification mechanism that is effective and protects privacy. We adapt the framework for deep speakers with high integrity accuracy based on the residual networks. Users can safely access home automation devices with SEASON. A thorough security analysis demonstrates that SEASON can successfully fend off various security threats. Additionally, the verifier's accuracy can be ensured.

Vinod Gujral *et al.* [19] suggested a frame-by-frame analysis of the voice to identify and authenticate the speaker based on its frequency and phase characteristics. We frequently encounter the issue of corrupt or synthetic speech, which enables unfettered access in various contexts. To put, anyone may corrupt the system. This results from a small dataset or a dataset with relatively few examples of synthetic speech.

Srinivas Parthasarathy *et al.* [20] experimented with a study of speaker verification performance with expressive speech. The effectiveness of emotion-based speaker verification systems is examined in this paper. They use the analytically helpful characteristics of arousal, valence, and persistent dominance rather than categorical categories with high inter-class heterogeneity, like happiness or rage. The speaker verification system in expressive speech can be predicted using estimated values for arousal, valence, and dominance.

Weight-based updates can be carried out using a complete network state history using recurrent training algorithms like Backpropagation through Time (BPTT) and Real-Time Recurrent Learning (RTRL). Generally, this process has some drawbacks: over time, their backpropagated error may be increased, and protecting this error from learning dependencies is more than a time step in length [21]. The majority of systems, however, use independent text modes with a Gaussian Mixture Model or i-vectors for speaker recognition. Although these techniques work, they can be avoided if a person has a voice recording of a recognized individual. The user must utter a unique passphrase to be recognized can resolve this. The system will first utilize voice recognition to verify the statement and then use speech authentication to identify the person, which will aid in double authentication [22].

3. METHODOLOGY

The proposed work has two phases: a training or registration phase and a verification or authentication phase. The many parts of the model are covered in great detail in the next section. Speech must be acquired, preprocessed, feature extracted, trained, and classified to recognize addresses using LSTM networks. Each step combines standard voice authentication system features with particular LSTM network properties. This work discusses creating a voice authentication system that directly processes the voice signal vectors using the LSTM architecture. Real-time voice authentication is used during the testing phase to record spoken sentences. Preprocessing and extraction, carried out on the test signal, have functions identical to those in the training phase.

The training data is run using the LSTM network, which is subsequently utilized to categorize the features.

The LSTM input layer receives speech functions as inputs, and a model corresponding to speech signals is built and stored in the word dictionary. The input layer, LSTM layer, connected layer, softmax layer, and output layer are the five layers that make up the LSTM network. After training using the collected characteristics, the LSTM network functions as a classifier, estimating the likely expression for categorization. The test dataset includes 11 frequently used terms in connection with gadgets. The deep convolutional neural network (DCNN) uses an input, output and two hidden layers. The characteristics are extracted from the audio-only portion in the same manner as the audio model and built into an improved deep convolutional neural network (IDCNN).

The produced model is an exact reproduction of the audio model, except there is an additional dense layer in place of the softmax layer. The generated feature map is fed into a DCNN with three dense layers. The batch normalization and dropout layers are placed after the first two thick layers in the deep neural network. The three earlier components are then combined as a final phase, resulting in a vector that combines the output vectors of the three earlier components. This is passed to a DCNN with three dense layers, a batch normalization layer, and a dropout layer. A step involved in the proposed voice authentication system is depicted in Figure 1.

3.1 Enrollment Phase

3.1.1 Pre-Processing

The collected voice data is currently being checked for errors. This is accomplished by analyzing data at various frequencies of various scales. Moreover, any clipping in the generated wavelets is examined. The discovered noise wavelets are then eliminated, and noise-free data is acquired. There are two approaches to eliminating noise from the audio data gathered. That is, applying the recursive least squares adaptive filtering approach and the threshold-based denoising method.

3.1.2 Feature Extraction Using VGG-16

Effective feature extraction techniques are used on the filtered signal to extract special features from each speaker's movement. 13 qualities are achieved with the VGG16. Nowadays, several CNN models have deeper architectures and excellent performance. On the other hand, deeper networks are challenging to train since they need a large amount of data and millions of parameters. For more realistic and generalized models, it is essential to include substantial and properly annotated data collection. Transfer learning techniques address this issue, where the model is initially trained and modified to handle the particular case.

VGG-16 network is applied for feature extraction. These operations serve as the input signal for the LSTM. VGG Net has 16 convolutional layers and 3×3 filters of three fully connected (FC) layers, and one convolutional layer stride is shown in Figure 2. The depth of the VGG-16 network is improved by including numerous tiny cores layered with filters, enabling it to extract more sophisticated functionality at a reduced cost. These functions are used as input signals for LSTM.

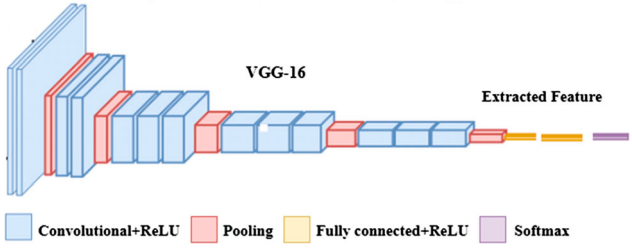

Figure 2: VGG16 Architecture

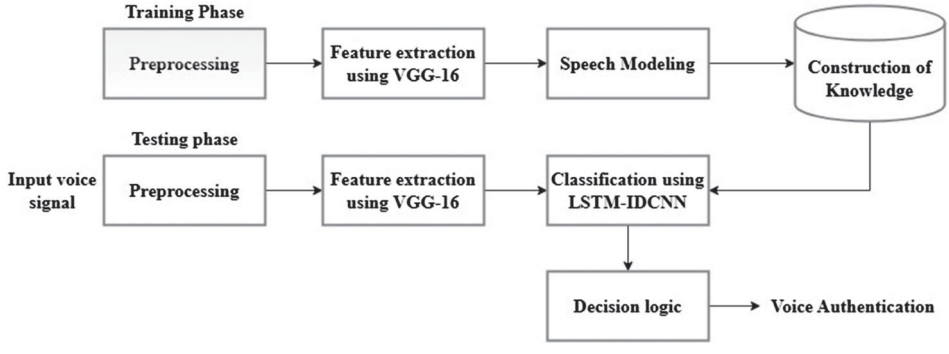

Figure 1: Steps involved in Proposed Voice Authentication system

The VGG-16 network contains many small stacked cores with filters that increase network depth, allowing for the extraction of more complex features at a lower cost.

3.2 Verification/Recognition Phase

3.2.1 Improved Deep Convolutional Neural Network (IDCNN)

The process flow diagram is shown in Figure 3 [23]. The IDCNN model focuses on architecture with a deeper level of detail that includes multilayer perceptron models with regularized learning techniques. Utilizing data analysis techniques for deep learning (DL) increases prediction accuracy while lowering skill requirements and the risk of human mistakes. Advanced deep learning prediction model and classification apply a more profound learning method based on deep multilayer perception, sophisticated and secure classification model with nonlinear functions and linear, regularization, descent, and binary sigmoid classification. The multilayer perceptron model with close learning of regularity is covered in greater detail by the IDCNN

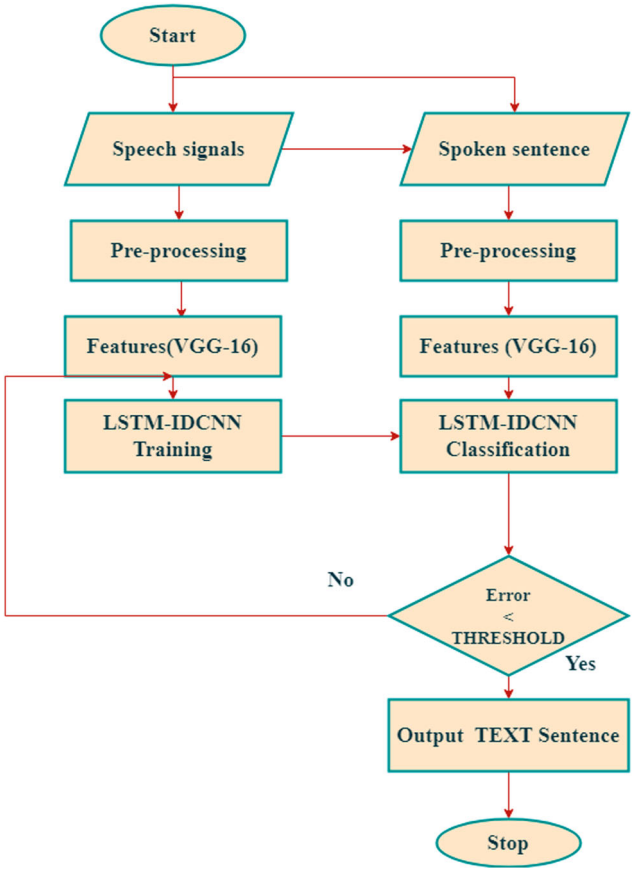

Figure 3: Flow chart for Voice Authentication Process

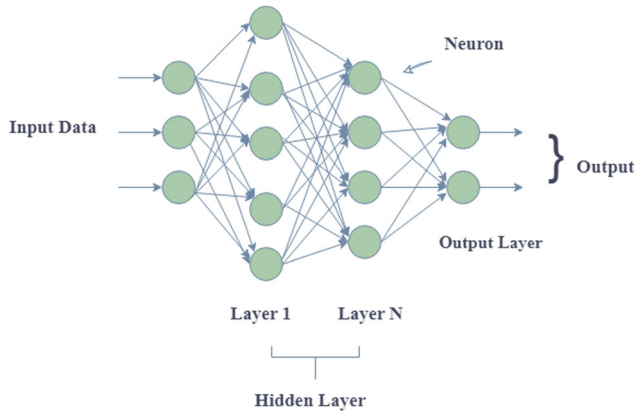

Figure 4: Deep convolutional neural networks

model [24]. The proposed IDCNN model is shown in the Figure 4.

3.2.2 Long Short-Term Memory (LSTM)

Cells make a typical LSTM network; these are memory blocks. The concealed and cell states are the two inherits of the subsequent cell. The drive's state is the primary data flow chain, and data can go through it unaltered. Data can be added to or deleted from the state of the cell through the sigmoid gates. Similar to a layer with various weights or a sequence of matrix operations, a gate. LSTMs are made to prevent long-term dependencies by employing gateways to manage stored procedures. LSTMs were chosen over standard RNNs because RNNs have short-term memory and cannot remember information from earlier steps.

Therefore, employ LSTMs to find a solution to this issue. The entire process is identical to a basic RNNS, except for the cell state and gate used in the LSTM. Data is kept in cell states during the processing of sequences. Using gates, information is added to or removed from the state of the cell. These gates are neural networks that choose the data for storage after decoding basic correlation structures. The input gate, the output gate, and the forget gate are the three gates that decide the network's hidden state [25]. The proposed LSTM architecture is depicted in Figure 5.

The information will be "lost" if the data is multiplied by 0, and it will be "retained" if multiplied by 1. As indicated in Figure 1, the output will be buried, denoted by, and the current state will beC_t if the current time step is t and the input value is x_t. Utilize the concealed state, H_{t-1}.,. The present state C C_{t-1} is the result of the forgotten state f_t and the prior storage state c. Whether the preceding memory cell value should be discarded is determined

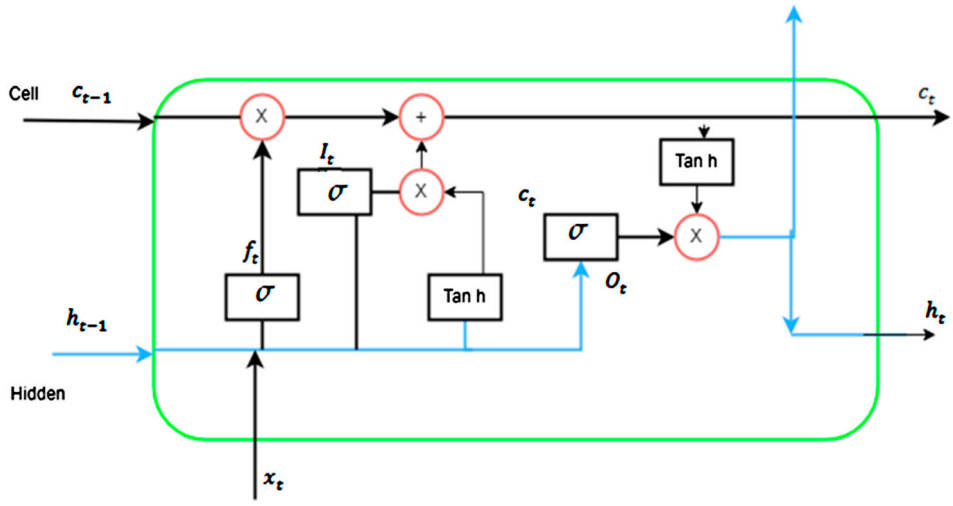

Figure 5: LSTM with internal structure

by the forget gate output. If the value is 0, it is set to be discarded; if not, it is kept. The exit gate establishes the next concealed state since it has an expected value. Use a variety of activation functions to activate a cell. The tanh function is assigned to the new state of the cell, which is the multiplication of the tanh output value and the sigmoid function, and the sigmoid function is applied to H_{t-1} and x_t to obtain the hidden state. The current time step's final hidden form is transmitted to the following time step [26,27].

The step by step procedure of the IDCNN-LSTM model is given below:

Step 1: Load dataset (1),(2) and (3)

Step 2: Removing noise using Hamming window.

Step 3: Make the spectral form of the speech signal frequency more Smooth using Pre-emphasis.

Step 4: Human hearing activity to detect frequencies using VGG-16.

Step 5: Digital representation of signals using Vector Quantization (VQ)

Step 6: Ruffle the dataset randomly and divide it into (k) groups using Cross-Validation, a(k-1)sub-instances were applied for training.

Step 7. The remaining data single sub-instance will be applied as the validation Data for testing.

Step 8. Classify instances using IDCNN-LSTM.

Step 9. Classifiers Evaluation.

Step 10. Best Classifier performance (Accuracy, Precision, Recall, F-measure, Error rate).

```
Algorithm for proposed work
Input: Voice dataset
Output: Best classifier performance
Begin
Best index = 0
If (Recognition Error Rate > 0) then
/* Start Find Best Distance word*/
    For each word of references template db
      Calculate distance with word and system output
    End For
    For each n
    /* n is the number of distance with word*/
    Search the best distance index(i)
    End For
    /* output correction process*/
    Output = word(i)
Else
    /* End of learning process (No correction needed)*/
End If
```

4. EXPERIMENTS

4.1 Dataset

Data is collected using various methods. We have collected recording data from three individuals (A, B, and C) in three 2-second audio clips, each with a microphone and a unique recording device, using a self-created online survey. The data set contains approximately five samples per individual. All audio clips in the speech authentication model must be in the same format. Audio files are converted to.wav format with a sampling rate of 16 kHz. Instead of using the recording's direct functions, perform a quantitative comparison of recording pairs from the same or different speakers using different messages.

Our dataset has the following features:

- Number of speakers: 3 (A, B, C)
- Intonations by speaker: normal, hoarse voice, whisper, nasal, fine voice,
- Number of different messages: 50

Dataset 1- If the speakers differ, pairs are created by comparing each recording of person A with each recording of person B. Following that, the first-class pair's sum across the entire dataset equals the previous value multiplied by different messages and the total speaker system: 2500 (*i.e.* 25*50*2).

Dataset 2 - If the speakers differ, pairs are created by comparing each recording of person A with each recording of person C. Following that, the first-class pair's sum across the entire dataset equals the previous value multiplied by different messages and the total speaker system: 2500 (*i.e.* 25∗50∗2).

Dataset 3- If the speakers differ, pairs are created by comparing each recording of person B with each recording of person C. Following that, the first-class pair's sum across the entire dataset equals the previous value multiplied by different messages and the total speaker system: 2500 (*i.e.* 25∗50∗2).

There are 2500 data samples in the entire data set. Three parts make up the adjusted amount of data. Mutual validation accounts for 20% of the total, testing accounts for 20%, and training accounts for 60%. A recording device was used to capture the first sound sample. An audio recording is transformed into an image of the sound using an image-based technique known as a spectrogram. Spectrograms are visual representations of the frequency spectrum, called the "voice impression." Spectrum experts in the two recordings highlight different words and phrases. Then, experts can look at a specific pattern of data in the image to determine how similar it is. Figure 6 depicts a spectrogram as an example.

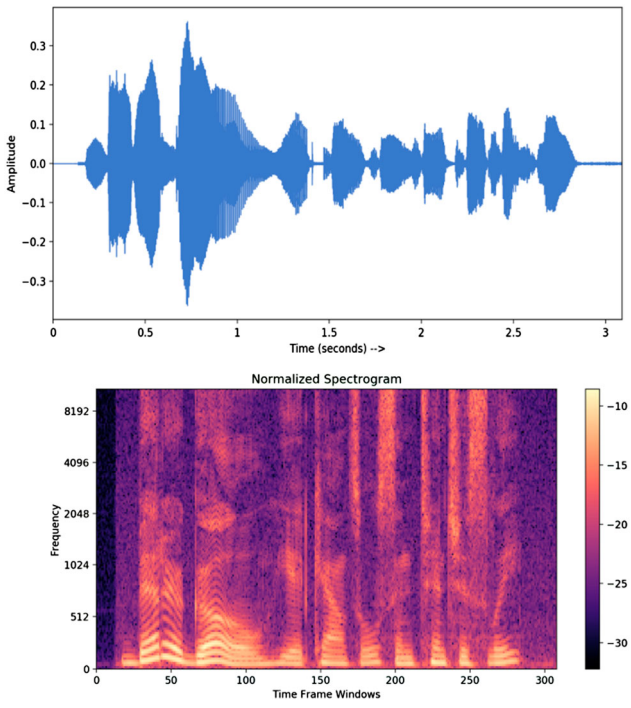

Figure 6: Audio file and Spectrum of Voice data

4.2 Implementation Details

This study employs a variety of grading scales, including accuracy scores, losses, classification reports, and confusion matrices. This section provides a detailed analysis of the model mentioned above's performance on both data sets across multiple rating scales. This section presents the suggested system's qualitative and quantitative examination. The software used by this system is 64-bit MATLAB 2021a. The objective classifications and segmentation comparisons are documented in terms of sensitivity, specificity, accuracy, and geometric mean (Gmean). Classification rates, which are numerical ratios, are used in the proposed IDCNN-LSTM hybrid improvement study. The amount of correctly identified objects is related to the total number of things utilized in the procedure.

$$Sensitivity\ (Sen) = TP/(TP + FN)$$

$$Specificity\ (Sp) = TN/(TN + FP)$$

$$Accuracy\ (ACC) = (TP + TN)/(TP + FN + TN + FP)$$

$$Geometric - mean = \sqrt{Sensitivity.Specificity}$$

$$Precision = TP/(TP + FP)$$

$$Recall = TP/(TP + FN)$$

$$F - measure = 2 * \frac{Precision * recall}{Precision + recall}$$

$$Error\ rate = 1 - [Accuracy]$$

4.3 Comparison Analysis

For comparison, we have utilized four benchmark models: SVM, ANN, NB, and LSTM. Figure 7 and Table 1 evaluate the outcomes of the suggested simulation results with those of other accepted techniques for the same dataset. These analyses reveal that the recommended strategy, which uses a CNN deep network score, yields higher simulation values from the same dataset than other tested techniques [4]. All accuracy, sensitivity, and specificity measurement information for all conventional models are included in Table 1. The accuracy of the suggested IDCNN-LSTM hybrid is increased by 98.83%.

Conventional LSTM comes second with 96.40%, and the SVM model performs the worst with an accuracy of

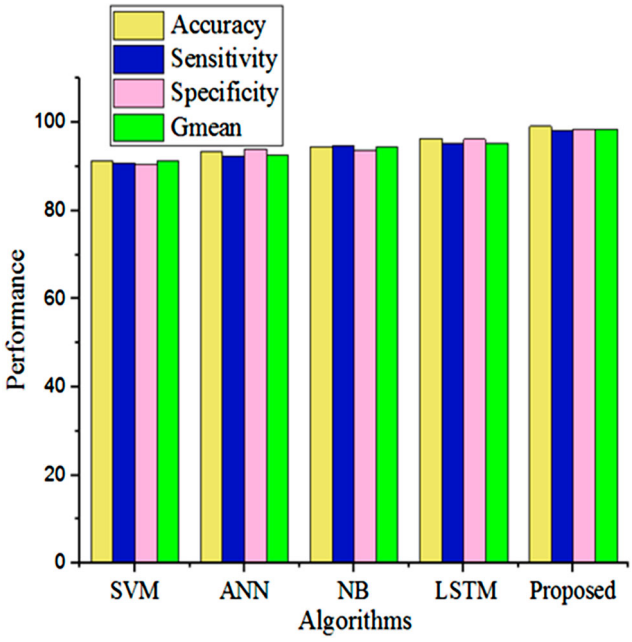

Figure 7: Comparison of proposed algorithm

Table 1: Comparison of proposed method with the literature works

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	G-mean (%)
SVM [28]	91.3	90.8	90.6	91.2
ANN [8]	93.5	92.4	93.1	92.6
NB [4]	94.3	94.6	93.8	94.5
LSTM [17]	96.4	95.3	96.2	95.3
Proposed	98.8	97.1	97.9	98.3

91.3%. The accuracy, sensitivity, specificity, and Gmean of the IDCNN-LSTM models are (98.8, 97.1, 97.9, and 98.3, respectively). The audio authentication model's accuracy and loss curve for the proposed algorithm is shown in Figure 8. The final experiments employ the proposed IDCNN-LSTM hybrid model, which performs admirably on training and test data sets.

Table 2: Performance Evaluation of the Proposed Method using dataset 1

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)	Error rate (%)
SVM [28]	85.8	84.9	85.1	84.8	11.8
ANN [8]	87.7	86.8	87.2	86.2	9.3
NB [4]	89.8	88.7	90.2	88.9	7.9
LSTM [17]	92.9	91.6	93.2	91.3	4.3
Proposed	94.7	98.1	97.8	96.5	1.7

The accuracy and loss of the LSTM model were perfect for training and test data. The final trials employ the proposed IDCNN-LSTM hybrid model, which performs brilliantly on training and testing data. The LSTM model outperforms in terms of accuracy and loss of training and testing data.

Figure 7 displays the outcomes from using the IDCNN-LSTM model to train and test the dataset. Other datasets produced similar results. As seen in Tables 2–4 and Figures 9–11, the proposed method classifier provides the highest accuracy, recall, precision, and f-measure for datasets 1, 2 and 3. The Naive Bayes classifier has the highest error rate, whereas the proposed method has the lowest error rate of the existing classifier. With this data set, the proposed classifier has the highest accuracy, recall, precision, and f-measure, while the SVM classifier has the lowest accuracy, recall, and f-number. In terms of error rate, the SVM classifier produces the highest error rate, while the proposed method yields the lowest error rate among other classifiers.

4.4 Comparative Analysis of Accuracy with Benchmark Algorithms

We have compared the proposed approach with multiple deep learning and machine learning state-of-the-art models. For comparison, we have utilized four

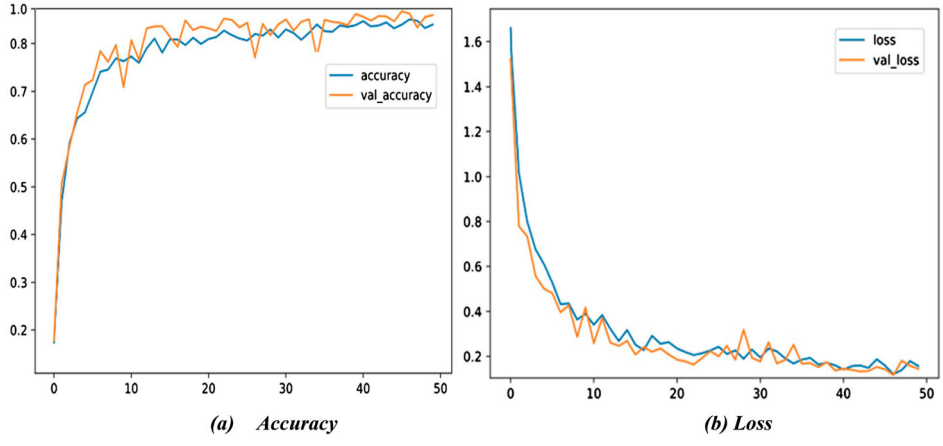

Figure 8: Accuracy and loss analyses

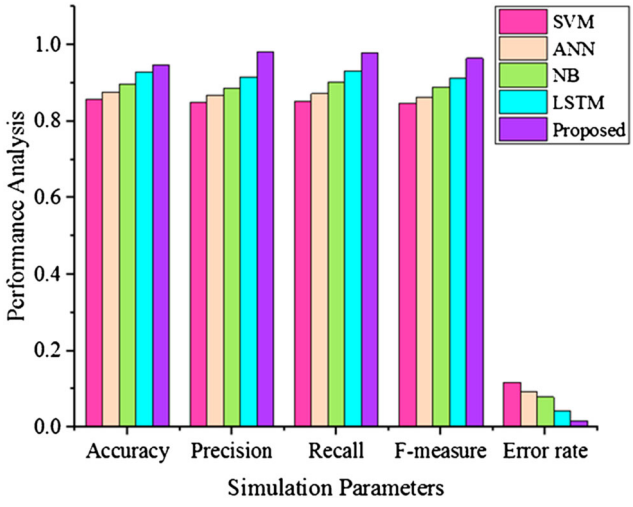

Figure 9: Performance Evaluation - dataset 1

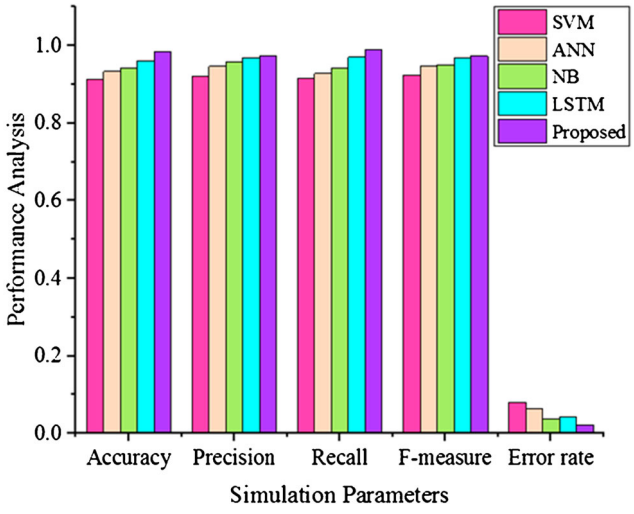

Figure 10: Performance Evaluation -dataset 2

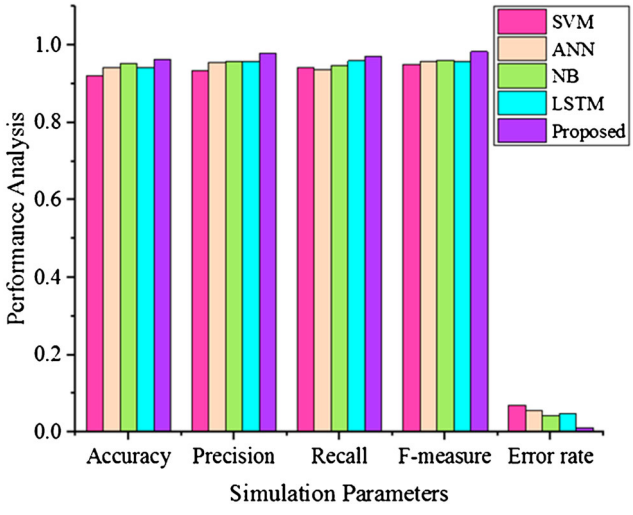

Figure 11: Performance Evaluation -dataset 3

Table 3: Performance Evaluation of the Proposed Method using dataset 2

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)	Error rate (%)
SVM [28]	91.4	92.1	91.5	92.4	8
ANN [8]	93.3	94.7	92.8	94.8	6.3
NB [4]	94.3	95.7	94.3	95.1	3.6
LSTM [17]	96.1	96.8	97.0	96.8	4.4
Proposed	98.3	97.5	98.9	97.3	2.1

Table 4: Performance Evaluation of the Proposed Method using dataset 3

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)	Error rate (%)
SVM [28]	92	93.5	94.3	95	7
ANN [8]	94.3	95.6	93.7	95.7	5.7
NB [4]	95.2	95.8	94.8	96.1	4.3
LSTM [17]	94.1	95.7	96.0	95.8	4.7
Proposed	96.3	97.9	97.2	98.3	1.1

Table 5: Comparative analysis of accuracy (%)

S.NO	Algorithms	Dataset 1	Dataset 2	Dataset 3
1	Decision Tree [29]	63.54	74.22	78.2
2	MLP Classifier[30]	78.38	86.66	82.5
3	NB [4]	89.86	94.3	88.37
4	CNN [9]	91.73	95.19	92.6
5	RESNET 50 [31]	93.16	97.06	95.2
6	IDCNN-LSTM	94.23	98.81	97.8

benchmark models: Decision Tree, MLP, NB, CNN and ResNet50. All these classifiers have been utilized and set to their base configuration. The model configuration of ResNet50 was set the same as VGG16 to keep an adequate comparison. The comparison of the accuracy of the benchmark and proposed models is shown in Table 5. It suggests the highest accuracies obtained by the proposed approaches concerning other conventional models.

4.5 Computational Complexity

Traditional voice authentication systems require many parameters and are computationally complex to implement. These methods are frequently used in "offline" environments where results are not produced in real-time. Traditionally, they have been designed to process voice recordings "offline," meaning that the system must wait for a period of time before determining the voice's identity. Because systems typically use millions of parameters, megabytes of memory and billions of algebraic calculations per recording is required. Voice authentication methods can be improved by reducing the number of parameters and calculations required, allowing them to perform better "real-time" speech processing. This research aims to reduce the computational complexity of the IDCNN-LSTM method so that it can produce results in less time. The preliminary findings of this study show

Table 6: Comparative analysis of the execution time (sec)

Classifier	Execution time (sec)
SVM [28]	365.90
ANN [8]	320.5
NB [4]	280.75
LSTM [17]	236.8
Proposed	170.23

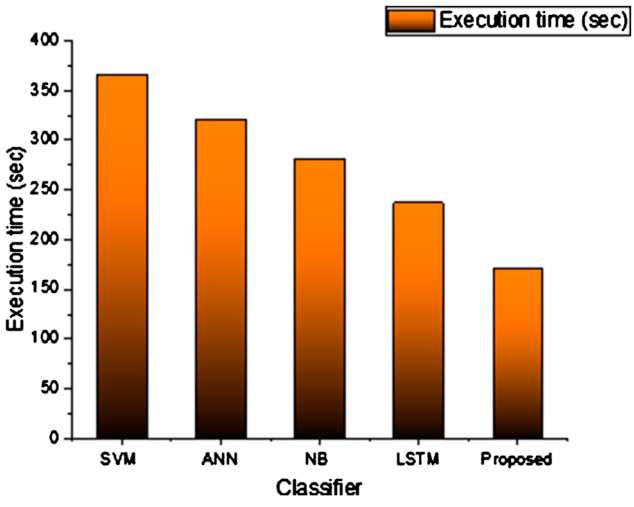

Figure 12: Execution time comparison

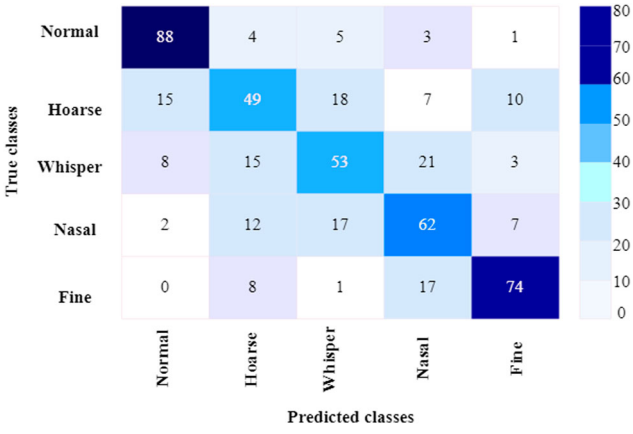

Figure 13: Confusion matrix of dataset-1

that it is possible to reduce the number of parameters while still maintaining voice authentication performance.

Table 6 and Figure 12 compare the execution times of the proposed methods. Existing algorithms such as SVM, ANN, NB, and LSTM achieved times of 365.90s, 320.5s, 280.75s, and 236.8s, respectively. Compared to the traditional algorithm, the proposed IDCNN-LSTM model takes 170.23 s to execute. Figures 13–15 depict these, with actual and predicted speech listed on the vertical and horizontal axes. From data set 1, the proposed IDCNN-LSTM model correctly identified 88% of normal speech, 49% of hoarse speech, 51% of whispery speech, 62% of nasal speech, and 74% of fine speech, as shown in Figure 13.

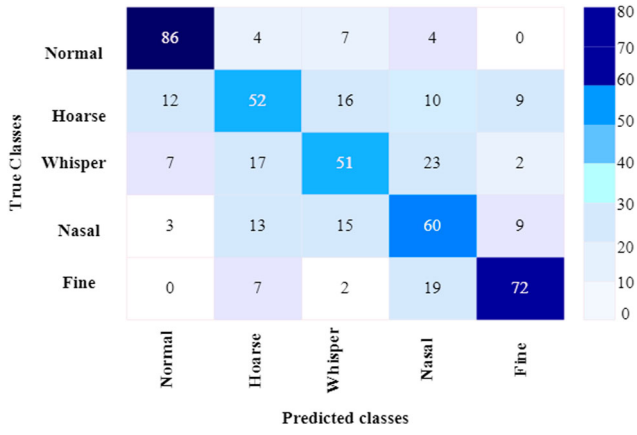

Figure 14: Confusion matrix of dataset-2

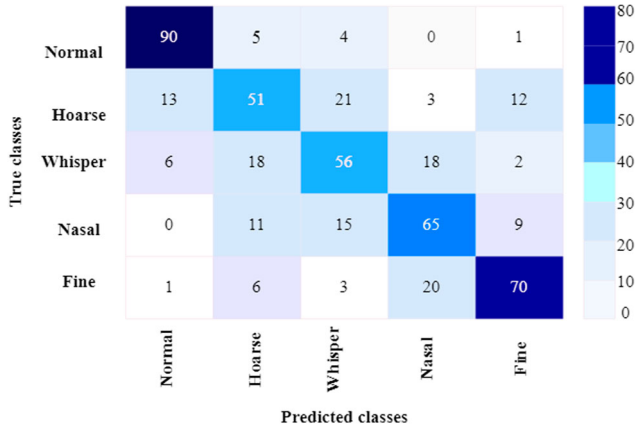

Figure 15: Confusion matrix of dataset-3

From data set 2, the proposed IDCNN-LSTM model correctly identified 86% of normal speech, 52% of hoarse speech, 51% of whispery speech, 60% of nasal speech, and 72% of fine speech, as shown in Figure 14. From data set 3, the proposed IDCNN-LSTM model correctly identified 90% of normal speech, 51% of hoarse speech, 56% of whispery speech, 65% of nasal speech, and 70% of fine speech, as shown in Figure 15.

4.6 Ablation Study

In this section, we conduct ablation experiments to test the efficacy of our proposed model. Table 7 contains all suggested datasets. Ablation studies show various architectures and authentication results that statistically represent the best models for the proposed task. Our ablation study evaluated different CNN architectures, chose the best one and began additional research.

Table 7: An ablation study of our proposed system configuration for 3 different datasets

Input	Architecture	Dataset-1	Dataset-2	Dataset-3
Audio Samples	CNN + Softmax	66.29%	68.32%	72.38%
	CNN + Softmax + FC	67.44%	68.76%	75.93%
	CNN + LSTM + Softmax	68.20%	70.48%	80.20%
	CNN + Softmax + FC + LSTM	71.37%	74.83%	82.12%
	ICNN + Softmax + LSTM	85.97%	87.78%	93.38%
	IDCNN + LSTM + Softmax + FC	93.940%	97.70%	96.89%

Table 7 displays the authentication results of various deep learning architectures using speech signals and improved deep convolutional neural network layers. In this study, we can see different results to evaluate our data, and we select the best models and process them. First, we validate the voice signal using only CNN and softmax. The results for this model were not convincing, so we added a fully connected layer, which slightly improved the results. Furthermore, we use a CNN to tune the global weights for long-term context dependencies for efficient voice authentication in long dialogue sequences. We use an LSTM network to rescale the global weights based on the learned features. This sequential learning model boosts authentication rates. We chose the IDCNN model for sequential learning because it performs well with partial data, but IDCNN with LSTM produces better results and is more appropriate for this task, as shown in Table 7.

5. CONCLUSION

The three main parts of this study are feature extraction, speech preprocessing, and machine learning classification. Audio preprocessing was a crucial component of the research as the recording used in our study was not recorded in a small area. This work suggests an improved hybrid acoustic model that takes advantage of improved DCNN and LSTM. It is proven that the IDCNN-LSTM architecture is reliable for voice recognition activity. All CNN models' performance is assessed using the performance metrics sensitivity, specificity, accuracy, and Gmean. The IDCNN-LSTM model, which has an overall accuracy of 98.83%, surpasses other machine learning algorithms. Compared to earlier methods, experimental data demonstrate that IDCNN-LSTM correctly classifies the speech signal. Accuracy, Sensitivity, Specificity, and Gmean are characteristics of IDCNN-LSTM models. These same models can be trained with various feature vectors in the future to assess their performance. To increase accuracy and decrease EER, deeper CNNs and LSTMs can also be used.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

REFERENCES

1. M. Kunz, K. Kasper, H. Reininger, M. Möbius, and J. Ohms. "Continuous speaker verification in realtime. Lecture Notes in Informatics (LNI)," in *Proceedings - Series of the Gesellschaft fur Informatik (GI)*. 2011, pp. 79-88.

2. M. N. Nachappa, A. M. Bojamma, C. N. Prasad, and M. Nithya, "Automatic speaker verification system," *Int. J. Res. Stud. in Comput. Sci. Eng.*, Vol. 1, no. 3, pp. 26–32, 2014.

3. A. Irum, and A. Salman, "Speaker verification using deep neural networks: A review," *Int. J. Mach. Learn. Comput*, Vol. 9, no. 1, pp. 20–5, 2019.

4. A. Ali, H. Abdullah, and M. Fadhil. "Voice recognition system using machine learning techniques." in *Materials Today: Proceedings*, 2021 doi:10.1016/j.matpr.2021.04.075.

5. R. M. Ghoniem, and K. Shaalan, "A novel Arabic text-independent speaker verification system based on fuzzy hidden Markov model," *Procedia Comput. Sci*, Vol. 117, pp. 274–86, 2017. ISSN 1877-0509, doi:10.1016/j.procs.2017.10.119.

6. B. Lamichanne, and K. C. Hari, "Performance analysis and recognition of speech using recurrent neural network," *Tech. J.*, Vol. 1, no. 1, pp. 87–95, 2019. doi:10.3126/tj.v1i1.27596.

7. A. Graves, D. Eck, N. Beringer, and J. Schmidhuber, "Biologically plausible speech recognition with LSTM Neural nets," *Lect. Notes Comput. Sci.*, Vol. 3141, 2004. doi:10.1007/978-3-540-27835-1_10.

8. N. Saxena, and D. Varshney, "Smart home security solutions using facial authentication and speaker recognition through artificial neural networks," *Int. J. Cognit. Comput. Eng*, Vol. 2, pp. 154–64, 2021. ISSN 2666-3074, doi:10.1016/j.ijcce.2021.10.001.

9. H. Salehghaffari. "Speaker verification using convolutional neural networks," 2018.

10. M. R. Ahmed, S. Islam, A. M. Islam, and S. Shatabda. "An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition," 2021.

11. S. Duraibi, F. T. Sheldon, and W. Alhamdani, "Voice biometric identity authentication model for IoT devices," *Int. J. Secur, Privacy Trust Manage*, Vol. 9, pp. 1–10, 2020. doi:10.5121/ijsptm.2020.9201.

12. J. Jung, H. Heo, I. Yang, S. Yoon, H.-J. Shim, and H. Yu. "D-Vector based speaker verification system using raw waveform CNN." 2018. doi:10.2991/anit-17.2018.21.

13. M. Bancroft, R. Lotfian, J. Hansen, and C. Busso. "Exploring the intersection between speaker verification and emotion recognition." in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2019, pp. 337-42. IEEE. doi:10.1109/ACIIW.2019.8925044.

14. T. Murugananthan, N. R. Nagarajan, and R. Balamurugan, "Biometric of speaker authentication using CNN," *Int J Future Gen Commun Netw*, Vol. 13, no. 1, pp. 1417–23, 2020.

15. H. Hertlein, A. Ariyaeeinia, Z. Jeffrey, and S. Ramalingam. "Effectiveness in the realisation of speaker authentication," in 2019 International Carnahan Conference on Security Technology (ICCST), 2019, pp. 1-5, doi:10.1109/CCST.2019.8888434.

16. F. Cheng, S.-L. Wang, and A. W.-C. Liew, "Visual speaker authentication with random prompt texts by a dual-task CNN framework," *Pattern Recognit.*, Vol. 83, pp. 340–52, 2018. ISSN 0031-3203, doi:10.1016/j.patcog.2018.06.005.

17. M. Dua, C. Jain, and S. Kumar, "LSTM and CNN based ensemble approach for spoof detection task in automatic speaker verification systems," *J. Ambient. Intell. Humaniz. Comput.*, Vol. 13, pp. 1–16, 2022. doi:10.1007/s12652-021-02960-0.

18. X. Gao, K. Li, W. Chen, W. Hu, Z. Zhang, and Q. Li. "Efficient and privacy-preserving speaker verification scheme for home automation devices," in 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2020, pp. 237-40, doi:10.1109/MIPR49039.2020.00056.

19. V. Gujral, J. Joshi, P. Medikonda, and N. Grover. "Advanced speech processing for speaker authentication in communication systems," in 2018 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), 2018, pp. 1-6, doi:10.1109/ANTS.2018.8710076.

20. S. Parthasarathy, C. Zhang, J. H. L. Hansen, and C. Busso. "A study of speaker verification performance with expressive speech," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 5540-4, doi:10.1109/ICASSP.2017.7953216.

21. S. Rudregowda, and S. Patilkulkarni, "Visual speech recognition for small scale dataset using VGG16 convolution neural network," *Multimed. Tools. Appl.*, Vol. 80, 2021. doi:10.1007/s11042-021-11119-0.

22. J. Jo, J. Kung, and Y. Lee, "Approximate LSTM computing for energy-efficient speech recognition," *Electronics. (Basel)*, Vol. 9, no. 12, pp. 2004, 2020. doi: 10.3390/electronics9122004.

23. V. Passricha, and R. Aggarwal, "A hybrid of deep CNN and bidirectional LSTM for automatic speech recognition," *J. Intell. Syst.*, Vol. 29, 2019. doi:10.1515/jisys-2018-0372.

24. L. N. Pondhu, and G. Kummari, "Performance analysis of machine learning algorithms for gender classification," *Proc Int Conf Inven Commun Comput Technol ICICCT*, 2018. doi: 10.1109/ICICCT.2018.8473192.

25. S. Rudregowda, S. Patilkulkarni, and S. B. Puneeth, "Combining audio and visual speech recognition using LSTM and deep convolutional neural network," *Int. J. Inf. Technol.*, 2022. doi: 10.1007/s41870-022-00907-y.

26. F. Ertam, "An effective gender recognition approach using voice data via deeper LSTM networks," *Appl. Acoust.*, Vol. 156, pp. 351–8, 2019. doi:10.1016/j.apacoust.2019.07.033.

27. E. J. Lok. "Toronto Emotional Speech Set (TESS)." in 2019. Available online: https://www.kaggle.com/ejlok1/toronto-emotionalspeech-set-tess (accessed on 16 December 2021).

28. P. Dhakal, P. Damacharla, A. Y. Javaid, and V. Devabhaktuni, "A near real-time automatic speaker recognition architecture for voice-based user interface," *Mach. Learn. Knowl. Extr*, Vol. 1, pp. 504–20, 2019. doi:10.3390/make1010031.

29. Y. Hu, D. Wu, and A. Nucci, "Fuzzy-Clustering-Based decision tree approach for large population speaker identification," *IEEE Trans. Audio, Speech, Lang Process*, Vol. 21, no. 4, pp. 762–74, 2013. doi:10.1109/TASL.2012.2234113.

30. T. B. Mokgonyane, T. J. Sefara, T. I. Modipa, and M. J. Manamela. "Automatic speaker recognition system based on optimised machine learning algorithms," in 2019 *IEEE AFRICON*, Accra, Ghana, 2019, pp. 1-7, doi:10.1109/AFRICON46755.2019.9133823.

31. R. Zheng, Y. Fang, and J. Dong, "Voice print recognition check-in system based on resnet," *Highlights Sci, Eng Technol*, Vol. 16, pp. 98–108, 2022. doi:10.54097/hset.v16i.2473.

AUTHORS

N Kaladharan received his under graduate in electrical and electronics engineering in 2004 and obtained the masters in computer science and engineering 2009 from Annamalai University. He is doing his research in speech processing. Currently, he is a lecturer of Department of Computer Engineering at Government Polytechnic College, Kottur - Theni. His research interests are speech processing, machine learning areas in the academic environments.

Corresponding author. Email: nkeeeau@gmail.com.

R Arunkumar received his under graduate in electronics and communication engineering from AVC College of Engineering, Tamilnadu in 2000 and obtained masters in computer science and engineering 2005 from Annamalai University. He received PhD in computer science and engineering from Annamalai University in 2016. He has published many technical papers in national and international conferences and journals. Currently, he is an associate professor of computer science and engineering, Faculty of Engineering and Technology, Annamalai University, Tamil Nadu, India. He is an associate editor and reviewer for several national and international journals. His research interests include image processing, machine learning, data and computer networks, information coding techniques, internet of things.

Email: arunkumar_an@yahoo.com