Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Deep Feature Extraction-based Speech and Speaker Recognition System using Heuristic Adopted Transformer Bidirectional Long Short Term Memory with Attention Mechanism

[1]Dr. Sukumar B S, [2]Dr. Sendamarai P, [3]Mrs. Kavya S, [4]Mrs. Pavithra S G, [5]Dr. Lakshmipathi M

[1]Associate Professor, Dept. of ECE, CBIT, Kolar, Karnataka, India 563101, sukumar.svm@gmail.com
[2]Professor, Department of ECE, Cambridge Institute of Technology North Campus, Bangalore Rural, Karnataka India - 562110 sendamarai.ece.nc@cambridge.edu.in
[3]Assistant Professor, Dept. of ECE, CBIT, Kolar, Karnataka, India 563101, ghanavigk@gmail.com
[4]Assistant Professor, Dept. of ECE, Bangalore Institute of Technology, Karnataka, India-560004 pavithrasg@bit-bangalore.edu.in
[5]Associate Professor, Department of ECE, Kuppam Engineering College, Kuppam, AP, India-517425 lakshmipathiece@gmail.com

ARTICLE INFO

Received: 05 Nov 2025
Revised: 20 Dec 2025
Accepted: 01 Jan 2026

ABSTRACT

Power normalization and Endpoint recognition can be essential to the effectiveness of both automated speaker authentication and speech detection. Conventional approaches for endpoint recognition and energy normalization frequently fall short in non-stationary settings. The systems have been employed in the majority of representation learning methodologies to learn and extract latent features from fixed length input. To address these issues, deep structure-based speech and speaker recognition systems are introduced to handle the difficulties in speech dataset variations. Initially, the required input is collected from the standard internet databases and then the desirable values are extracted from the collected input that includes spectral features like spectral contrast, flux, spectral centroid, spectral flatness and spectral bandwidth, and cepstral features like Mel Frequency Cepstral Coefficient (MFCC) and Linear Predictive Coding Coefficient (LPCC) and finally the deep attributes are extracted with the support of Autoencoder network. Secondly, fused weighted parameter selection is carried out via the newly developed Hybridization of Bonobo with the Dandelion Optimization Algorithm (HBDOA). Thirdly, speaker recognition and speech recognition is carried out, where the speech and speaker recognition is done via Transformer Bidirectional Long Short Term Memory with Attention Mechanism (TransBiLSTM-AM). Here, the values within the Trans-BiLSTM are optimized using the implemented HBDOA. Finally, the implementation results are analyzed over distinct existing speech and speaker recognition systems over developed speaker and speech recognition model's performance.

Keywords: Speech and Speaker Recognition System; Hybridization of Bonobo with Dandelion Optimization Algorithm; Optimal Fused Weighted Feature Selection; Transformer Bidirectional Long Short Term Memory with Attention Mechanism

1. Introduction

The process of translating speech signals into text transcription is called speech recognition. The field of speaker and speech recognition has advanced significantly in recent years [1]. The evolution from isolated word to continuous voice recognition, from short vocabulary to broad vocabulary recognition

and from hardware recognizer to software recognizer to recognition are all examples of the progress [2]. Work on speech recognition is connected to speaker recognition. The emphasis is on identifying the speaker rather than figuring out what was said [3]. If a specific speaker was the one who made the speech is known as verification (Speaker recognition), and determining a person's identity from a list of recognized speakers is known as speech detection (Speaker verification) [4]. The most common type of speaker recognition is still not particularly accurate for huge speaker populations, but it is also be carried out on a workstation if the user's words are limited and the speech quality cannot vary too greatly [5]. Both speaker recognition and voice recognition have seen the majority of these advancements [6]. The availability of speech and voice recognition systems for computer communication has grown in popularity as computing power has increased. Many technologies, including personal digital assistants, search engines and mobile communication, extensively use speech-to-text or text-to-speech systems.

Nowadays, voice recognition is a challenging task. Since it is difficult to handle differences in systems and datasets with various languages and adapt to new languages [7]. The voice recognition algorithms must accommodate dataset variances like gender and accents [8]. There are a number of distorted factors that affect speech signals, like background noise. Deep learning methods achieved greater success in several applications like sequence alignment and image processing [9]. The deep learning method with optimization can overcome these difficulties [10]. Speech recognition systems became increasingly successful and human-computer interactions increased with the advent of deep learning technology [11]. All speech recognition applications, including personal digital assistance, mobile communication and voice search, are seeing great success and increased interest in deep learning-based systems [12]. They require enormous volumes of data to provide the best and better quality of service. Additionally, they have a lot of data on a certain subject, preventing us from using it to train supervised learning strategies [13].

The speech vision uses visual data augmentation and transfer learning techniques to dysarthric acoustic inputs to solve the problems of scarcity [14]. The accent and gender approaches are used in the speech recognition system. These techniques improved the performance and decreased the Label Error Rate (LER) [15]. The most popular architectures for voice recognition use deep learning methods [16]. A number of application domains are used for handwriting recognition, speech recognition, time series analysis, grammar learning and prediction systems. RNNs have the ability to handle sequential data and produce cutting-edge results. RNNs have been used in speech recognition for keyword spotting, voice activity identification, speech emotion recognition and speech improvement [17]. The RNN method is suffered from vanishing gradient problems. Therefore, the developed speech and speaker recognition system effectively identifies the appropriate person through the speech.

The objectives of the suggested speech and speaker recognition system utilizing deep learning are described below.

- To design a speech and speaker detection system for predicting the appropriate person through the speech.
- To design an optimal weighted feature selection fusion stage using HBDOA optimization to maximize the correlation coefficient by optimizing the weights for improving the recognition performance.
- To implement an efficient HBDOA for weight optimization and select the features in optimal weighted feature selection and optimize the variables like hidden neuron count, count of activation function, and epochs from the detection phase for enhancing the effectiveness of the model.
- To develop TransBiLSTM-AM-related detection for effectively identifying the person through their voice using HBDOA strategy with parameter optimization.

- To evaluate the implemented speech and speaker recognition performance over heuristic algorithms and several traditional methods.

The detailed summarization of the designed speech and speaker recognition system is presented in the below phases. The conventional detection system's parameters and challenges are provided in section II. The dataset description, proposed system and suggested strategy structure are provided in phase III. Phase IV explains the optimal weighted feature selection and fusion description. The prediction techniques details are specified in section V. Section VI provides the experimental setup of the investigated system and outcomes. The conclusion of the developed speech and speaker recognition system is specified in section VII.

2. Literature survey

2.1 Related Works

In 2021, Dokuz and Tufekci *et al.* [18] have designed a efficient technique-based speaker and speech detection system. The effective gradient optimization was used in this system. The deep learning-related speech detection model performance was improved by using four techniques and these techniques were used to extract the important features. The developed model adjusted the accent and gender using effective techniques in the dataset. The studies showed that the proposed strategies had given best effectiveness than traditional selection strategy.

In 2019, Lee *et al.* [19] have investigated a speech prediction system utilizing techniques. The best information was selected from a multi-condition dataset using mask-estimator-based neural networks. The developed speech detection model had given a high signal ratio due to the comparison of existing models. It decreased the insertion errors and allowed the deep acoustic and language models to better exploit the rich context information. The suggested model better than recently used approach with the help of reducing the error rates.

In 2018, Kumar *et al.* [20] have investigated a speech prediction model employing effective methods. The suggested model used an effective genetic strategy with an advanced method for feature extraction to maximize the selected features in the dataset. The initial phase was feature optimization with a genetic algorithm and the third and final phase was feature extraction with an effective deep learning method. Finally, a real-world situation was used to validate and evaluate the suggested work paradigm. The experimental outcomes showed that a high recall rate, accuracy, specificity, sensitivity, and precision rate was attained using the developed method compared to existing work.

In 2021, Reza *et al.* [21] have designed a framework for detecting speech using an advanced deep learning strategy with the automatic concept. Automatic speech detection for people contains many difficulties, such as speech vision and dysarthric-specific. The speech vision was used to recognize the shapes of the words produced by people with dysarthria. The revolutionary method for dysarthric was to extract speech features visually. These features were used to remove the phoneme-related difficulties. The speech vision used the augmented approaches of visual information to extract the dysarthric acoustic features for solving the data shortage issue. The cutting-edge methods of speech recognition were given less recognition accuracy compared to the developed system. The suggested speech recognition system showed greater effectiveness.

In 2017, Misra *et al.*[22] have proposed a multichannel-based target speaker identification using neural network architecture. This model used a deep learning framework to combine multichannel augmentation with acoustic modeling. The suggested model introduced a neural architecture in the network's first layer and it carried out multichannel filtering. The model demonstrated that it was given information of the real speaker directions in terms of performance than other models. The model showed how providing the initial layer to isolate the effective filtering might enhance performance. An adaptive variation that adjusted the spatial filter coefficients based on the inputs from the previous time frames was introduced in the first layer. It was shown that these methods

might be used more effectively in the frequency domain. These neural networks improved the word error rate more than a conventional beamforming-based framework.

In 2020, Devi *et al.* [23] have recommended a effective deep model for speech recognition. The effective method was extensively employed to select the signal features. These selected values were used to produce their input samples, and their size decreased using a feature map. Recognition was carried out using Bayesian Regularization. The 10 persons were used to train and verify the network. Performance estimation and prediction rates in compared among conventional models were used to validate the suggested technique.

In 2020, Kharroubi and Hourri *et al.* [24] have offered a speech recognition system using a unified approach based on deep learning to address effectiveness mortification problems caused by a mismatch between several conditions due to various inter-speaker variability. The source and target were the homogenous domains involved in the initial solution. The most important finding was that the transfer learning outperformed than existing algorithms. The testing results showed that the target speaker was provided very efficiently than previous models.

In 2022, Yang *et al.* [25] have investigated a advanced technique for the recognition of speech. The proposed algorithm utilizes high level parameters. The files of RGB images were used to to recognize emotions. The deep learning approach was used for the implemented system. The public available data set was used for the experimental analysis. This developed speaker and speech detection had given betterprediction rates.

2.2 Problem statement

There are still obstacles in speech recognition because of the hurdle of using methods and distorting factors, which are microphone quality, environmental factors and background noise. Due to the interference of speakers, background noises and room reverberation the target speech signals get corrupted, which results in low performance. Hence, deep learning-related speech prediction systems are offered to solve these disadvantages. The features and disadvantages of the deep learning-based speech recognition systems are specified in Table I. RNN and LSTM [18] have the potential to adapt new languages and also it can handle sequential data due to their high capability. Yet, they are sensitive to different random weight initializations and it is affected by environmental factors, background noise and unclear microphone quality. CGMM [19] helps to enhance scalability and robustness. In addition, it is used to reduce the noise efficiently. But, it degrades the efficiency due to the corruption of target speech signals. For instance, the shape of the cluster is undetermined. MFCC [20] smoothened the frequency response of the vocal tract and played a major role in error reduction. But, it is highly sensitive to noise due to its dependence on the spectral form and it is not robust enough in noisy environments. ASR [21] acts as a good intermediary for dysarthric persons. In addition, it improves the quality of life for dysarthric individuals. Though, they have poor recognition of dysarthric speech for severe dysarthria and do not organize the speech orderly. Neural network architecture [22] helps in the improvement of the relative word error rate. Moreover, it provides high specificity, precision and recall. But, they require a lot of data to train the classifiers, which is more computationally expensive than traditional algorithms. ANN [23] helps in the storage of information in the entire network. But this network requires lots of computational power and it requires lots of data for training. DNN [24] has multiple layers to learn complex features and they are used to learn the primary framework of information input data vectors. Though, it takes more time for classifier training and it is expensive to forecast the speech signals. DCNN [25] has a reduced number of parameters that helps to save memory. Moreover, the cost of computing is minimized due to weight sharing. Furthermore, they do not find the orientation of the things and Class imbalances are one of the major drawbacks. Therefore, these disadvantages helped to implement a speech and speaker recognition system.

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/ **Research Article**

Table 1 . Features and disadvantages of existing Deep learning Techniques for Speech and speaker recognition system

Author [citation]	Methodology	Features	Disadvantages
Dokuz and Tufekci [18]	RNN and LSTM	They have the potential to adapt to new languages. It has handled the sequential data due to its high capability.	They are sensitive to different random weight initializations It is affected by background noise, unclear microphone quality, and environmental factors.
Lee *et al.* [19]	CGMM	It helps to enhance scalability and robustness. It is used to reduce the noise efficiently.	It degrades efficiency due to the corruption of target speech signals. The shape of the cluster is undetermined.
Kumar *et al.* [20]	MFCC	The frequency response of the vocal tract is relatively smooth. It plays a major role in error reduction.	It is highly sensitive to noise due to its dependence on the spectral form. It is not robust enough in noisy environments.
Reza *et al.* [21]	ASR	They act as good intermediaries for dysarthric persons. It improves the quality of life of dysarthric individuals.	Poor recognition of dysarthric speech for severe dysarthria. It does not organize the speech orderly.
Misra *et al.*[22]	Neural network architecture.	Improvement in relative word error rate. It provides high specificity, precision and recall.	They require a lot of data to train the classifiers. It is more computationally expensive than traditional algorithms.
Devi *et al.* [23]	ANN	It helps in the storage of information in the entire network. It does not avoid the complex structures.	This network requires lots of computational power. It requires lots of information during the implementation period.
Kharroubi and Hourri *et al* [24]	DNN	It has multiple layers to learn complex features. They are employed to detecting framework of the input information vectors.	It takes more time for classifier training. It is extremely expensive to recognize speech signals.
Yang *et al.* [25]	DCNN	The reduced number of parameters helps to save the memory. The cost of computing is minimized due to weight sharing.	They do not detect the location of the things. Class imbalances are one of the major drawbacks.

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

3. A Primary Architectural Description of Proposed Network Intrusion Detection System using Deep Learning with Attention Mechanism

3.1 Structural View of Implemented Model

The existing speech and speaker recognition systems suffered from many challenges like imitation or mimicry, low-quality voice, attackers and fault tolerance. The impostor trying to impersonate a subject included in the framework to get access to the framework using an external record is known as imitation or mimicry. In the current work, speaker recognition faces a significant obstacle. The fact that language models need to be trained takes a lot of time and knowledge, which might be a problem for speech recognition. The attacker impersonates the target by using their fake voice. As a result, it struggles to understand the discourse. The conventional systems struggle to distinguish speech because of the poor quality of voice samples. The speaker's voice can alter the most due to surrounding noise, health issues, mood swings, prolonged use of digital and analogy microphones, and other factors. As a result of being used for a long time, hardware and software may no longer be able to support an algorithm that works wonderfully for speaker recognition, and vice versa may also occur. Voice technology also has privacy concerns in other industries, such as keeping financial and banking data secure or just hiding some information from ears that aren't supposed to hear it. Hence, the developed speech and speaker detection system are used to predict the appropriate person through the speech of the individual. The diagrammatical demonstration of the speech and speaker recognition system is displayed in Fig. 1.

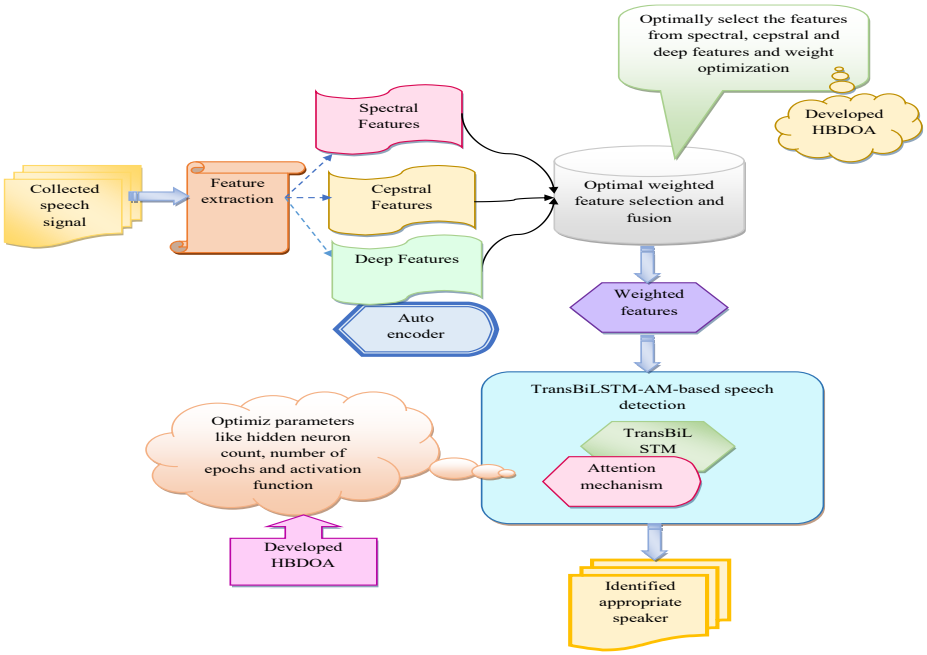


Figure 1 .Structural demonstration of speech and speaker recognition system using deep learning

The newly developed speech and speaker recognition system is used to identify the person through their voice with higher accuracy. The speech and speaker data is collected from the internet. The initial step is to features are extracted from collected data. The spectral extracted features like peak amplitude, spectral flux, spectral centroid, spectral density, standard deviation, spectral roll-off, entropy, zero crossing rate, total harmonic distortion and root mean square succeeding values are extracted. The cepstral features like MFCC and LPCC are extracted. The deep best values are removed utilizing the autoencoder technique. These removed variables are fed into the fused weighted feature selection section. Then, the suggested HBDOA algorithm is used to optimally select the best features from spectral, cepstral and deep features. Also, weight optimization is performed to improve the

correlation coefficient. The selected variables are concatenated with optimized weights to get fused weighted features. Then, the weighted features are fed into the prediction. The person identification is performed using the TransBiLSTM-AM approach. The suggested HBDOA is employed to optimize the parameters like the hidden neuron count, count of activation function and epochs to improve the accuracy and decrease the FDR, FPR, and FNR. Finally, it identifies and verifies the person through their voice effectively. The suggested speech and speaker prediction system performance is compared among conventional models with the help of different effectiveness metrics and it is given high accuracy.

3.2 Speech and Speaker Recognition Dataset

Dataset-1 (Tensor Flow Speech dataset): The input speech signal is downloaded from the below link "https://www.kaggle.com/competitions/tensorflow-speech-recognition-challenge/data" Access date: 2023-08-07. It contained 4 folders in a .txt format. The audio files are included in this dataset. It is one second video clip. The files contained several columns that named with down, left, yes, no, up, stop, go, right, on, off. The folder named like test.7z, train.7z and _background_noise_ and it contains the video comments in this dataset.

Dataset-2 (AudioMNIST): The input speech signal is downloaded from the below link" https://github.com/soerenab/AudioMNIST" Access date: 2023-08-07". It contains sixty person audio files and a total 30000 audio files are included in this dataset. One folder holds the total audio files. The "audioMNIST_meta.txt" folder is included in this dataset. It is the collection of information about the speaker person's age and gender information.

Hence, the total input data is indicated by B. The collected speech signal is noted by S_{hd}^{B}, where $b = 1,2,\cdots, B$.

3.3 Designed HBDOA

The suggested HBDOA strategy is used to enhance the effectiveness using the optimization of weights and parameters. The suggested HBDOA strategy is utilized to optimize the values like features from spectral features, features from cepstral features, and features from deep features and weight optimization is performed to enhance the correlation coefficient in the fused weighted feature selection phase. Also, the implemented HBDOA is used to optimize the parameters like hidden neuron, number of epochs and activation function from Trans-BiLSTM to enhance the accuracy and decreases the FDR, FNR and FPR. The BO algorithm easily solves complex problems and it does not stuck in the local minima. But, it is time consuming and the computational power is high. The DA optimization is effectively reducing the computation complexity issue and it is easy to implement. Yet, it is given poor accuracy in the prediction section. The cost of the implementation is high. Hence, the suggested HBDOA strategy is employed to resolve the above problems. In the developed algorithm, the final position is noted by Y_j and it is determined with the new formula by considering the positions obtained from BO and DA. The adaptive concept of the newly upgraded final position is given in Eq. (1).

$$Y_j = \left(\frac{\left(\left(Ky_i + KT_i \right) / 2 \right) + bestfit}{\sqrt{\left(bestfit \wedge 2 + wortfit \wedge 2 \right)}} \right) \tag{1}$$

Here, the best fitness denoted as $bestfit$ and the value $wortfit$ is the worst solution. The term Ky_i is the position of BO algorithm and the parameter KT_i is the position of DA optimization. The implemented HBDOA is employed to effectively solve complex optimization issues. It is employed to decrease the vanishing gradient issue.

BO [26]: The Bo algorithm is implemented based on the bonobo's social activities and behaviour. It is the population based strategy. The bonobo's fitness and population values are validated in the BO

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

algorithm. The term O is the random value, selected in the interval of $[0.0,1.0]$. The temporary factor size is validated employing Eq. (2).

$$utht_{\max} = Max(2, utht_{fact} \times O) \tag{2}$$

Here, the term $utht_{fact}$ is the temporary factor size. The bonobo's minimum parameter is set to 2. The term q_q is the initial parameter and it is set to 0.5. The new bonobo is measured using Eq. (3).

$$\begin{aligned} NW_k = BO_k^j + s_1 \times tdbc \times \left(\beta_k^{BO} - BO_k^j\right) + \\ (1 - s_1) \times tdbc \times gmbh \times \left(BO_k^j - BO_k^Q\right) \end{aligned} \tag{3}$$

Here, the parameter NW_k is the new bonobo. The intermediate values are indicated by α_1 and α_2, respectively. These parameters are measured using Eq. (4) and Eq. (5), respectively.

$$\alpha_1 = f^{\left(s_4^2 + s_4 - \frac{2}{s_4}\right)} \tag{4}$$

$$\alpha_2 = f^{\left(-s_4^2 + 2s_4 - \frac{2}{s_4}\right)} \tag{5}$$

These parameters helped to find the new bonobo. The lower and upper limits are denoted by Min_k and Max_k, respectively. The random variable s_4 is selected in the range of $[0.0,1.0]$. The new bonobo is calculated by Eq. (6), Eq. (7), respectively.

$$NW_k = BO_k^j + \alpha_1 \times \left(Max_k - BO_k^j\right) \tag{6}$$

$$NW_k = BO_k^j - \alpha_2 \times \left(BO_k^j - Min_k\right) \tag{7}$$

The terms $NW1_k$ and $NW2_k$ are validated by Eq. (8), Eq. (9), respectively.

$$NW1_k = BO_k^j + \alpha_1 \times \left(BO_k^j - Min_k\right) \tag{8}$$

$$NW2_k = BO_k^j + \alpha_2 \times \left(Max_k - BO_k^j\right) \tag{9}$$

The new parameter is upgraded related on the consortship mating technique and it is provided in Eq. (10).

$$NW_k = BO_k^j + gmbh \times \alpha_2 \times f^{-s5}\left(BO_k^j - BO_k^Q\right) \tag{10}$$

The random values are noted by s_5 and s_6, respectively. The variable q_q is the directional probability. The negative stage is indicated by OQ. The term $utht_{factInt}$ is measured using Eq. (11).

$$utht_{factInt} = 0.5 \times utht_{factMax} \tag{11}$$

Here, the term $utht_{factMax}$ is the maximum parameter of factor values. The initial parameter of probability value is set to 0.5.

DA [27]: it is developed related on the dandelion's sowing process. The sowing behaviour of dandelions on earth is separated into two categories. The term z is specified in Eq. (12).

$$z = \min g(y) \tag{12}$$

The term y is the optimal value. The dandelion seeds are spread on the dandelion's surrounding place. Hence, the DA algorithm is only focused on the dandelions nearest place. The term N_j is measured using Eq. (13).

$$N_j = \begin{cases} \max \times \dfrac{g_{\max} - g(y_j) + \varphi}{g_{\max} - g_{\min} + \varphi} & N_j > \min \\ \min & N_j \le \min \end{cases} \tag{13}$$

Here, the term $\max$ is the seed's maximum count. The seed's minimum count is indicated by $\min$. The machine epsilon parameter is represented by φ. The amounts of seeds are indicated by N_j. The core dandelion fitness function is calculated using Eq. (14).

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

$$Y_{DE} = \min g\left(_j y\right) \tag{14}$$

Here, the value Y_{DE} is the best solution of core dandelion. A suitable dandelion grows faster than an unsuitable dandelion place. The assistant dandelion's radius is calculated using Eq. (15).

$$S_j(u) = \begin{cases} VC - MC & u = 1 \\ x \times S_j(u-1) + \left(\left\|Y_{DE}\right\|_\varsigma - \left\|y_j\right\|_\varsigma\right) otherwise \end{cases} \tag{15}$$

Here, the weight is measured using Eq. (16).

$$x = 1 - \frac{Gf}{Gf_{\max}} \tag{16}$$

Here, the term Gf is the present fitness value. The fitness function's maximum count is indicated by $Gf_{\max}$. The core dandelion's radius is calculated using Eq. (17).

$$S_{DE}(u) = \begin{cases} VC - MC & u = 1 \\ S_{DE}(u-1) \times s & b = 1 \\ S_{DE}(u-1) \times f & b \neq 1 \end{cases} \tag{17}$$

Here, the term $S_{DE}(u)$ is the radius of the dandelion. The fading factor is indicated by s and the growth factor is noted by f. The location is indicated by Y_j. The growth rate is measured using Eq. (18).

$$b = \frac{g_{DE}(u) + \varphi}{g_{DE}(u-1) + \varphi} \tag{18}$$

Here, the term b is the epsilon value and it is set to 0. The best solution is used to enhance the speed of the convergence rate. The suggested HBDOA algorithm pseudo-code is shown in Algorithm 1. The flowchart of investigated HBDOA is specified in Fig. 2.

Algorithm 1: Designed HBDOA
Load the population of both BO and DA
Generate the parameters
Determine the best optimal solution with BO and DA
While $(v < v\max)$
For $(j = 1 to\ nPop)$
Update the first location Ky_i using BO
Update the second location KT_i using DA
Update the final position Y_j using the adaptive concept
End For
End while
Return the best fitness solution
End

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

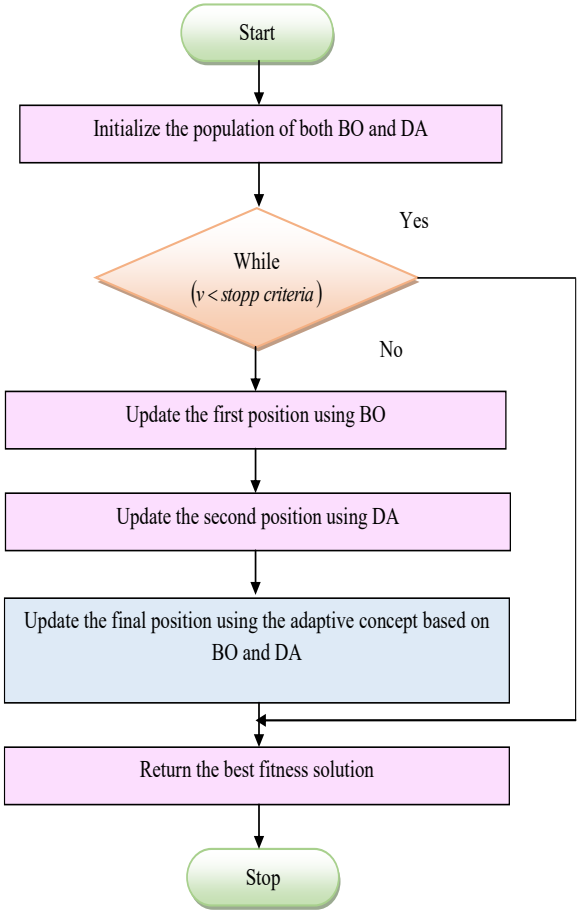


Figure 2 .Flowchart of investigated HBDOA

4. Spectral, Ceptral and Deep Feature Extraction and Optimal Weighted Feature Fusion using Proposed Optimization Strategy

4.1 Spectral Feature Extraction

The accumulated speech signal S_{hd}^{B} is inset into the spectral feature extirpation. Spectral feature extirpation is a method of extracting non informative and informative features using automated spectral analysis. The extraction of spectral features explanations are given below.

The Spectral centroid calculates the amplitude of the centre of the signal among frequencies. The Spectral centroid is calculated using Eq. (19).

$$D = \frac{\Sigma_{o=0}^{O-1} g(o) y(o)}{\Sigma_{o=0}^{O-1} y(o)} \tag{19}$$

Here, the variable $g(o)$ is the frequency and the term $y(o)$ is the frame's amplitude. The entropy is the randomness or impurity in the signal data. The entropy is calculated by Eq. (20).

$$\Delta X \infty \frac{\Delta I(z)}{\Delta X} X^{U} X = \left[J - \kappa(v) v^{U} X \right] \tag{20}$$

Here, the parameter $\kappa(v)$ is the density value. The speech segment size is denoted by z . The signal is indicated by v . The difference between signal and adjacent frame information is called spectral flux. The Spectral flux is calculated using Eq. (21).

$$Gm_{(j,j-1)} = \sum_{l=1}^{Xg_m} \left(FO_j(l) - FO_{j-1}(l) \right)^2 \tag{21}$$

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

The term $FO_j(l)$ is the normalized coefficient. The frame is indicated by j. The positive to zero and negative to positive signal transition rates measure is known as zero crossing rates. The zero crossing rates are calculated using Eq. (22).

$$Z = \frac{1}{U-1}\sum_{u=1}^{U-1} 1_{S<0}\left(t_u t_{u-1}\right) \tag{22}$$

Here, the RMSE calculates the difference between predicted values and parameters of the developed system. The RMSE formula is given in Eq. (23).

$$RMSE(y) = \sqrt{\frac{1}{O}\sum_{o}|y(o)|^2} \tag{23}$$

Here, the term $\sum_{o}|y(o)|^2$ is the mean value of the amplitude. The relation of signal data and mean is called standard deviation and it is given in Eq. (24).

$$W_k = \sqrt{\frac{\sum_{(k-1)\eta+1}^{k\eta}\left(a_j - \bar{a}_j\right)}{\eta - 1}} \tag{24}$$

Here, the signal size is noted by η. The parameter indicates the a_j time series. The harmonic distortion is the calculation of total voltage harmonics in the signal data. The total harmonic distortion variable is validated using Eq. (25).

$$T = \frac{\sqrt{\sum_{o=2}^{\infty} W_{no}^2}}{W_{fd\_sp}} \tag{25}$$

Here, the term W_{no}^2 is the voltage of the signal and the term W_{fd_sp} is the frequency. These features are concatenated to get final features and it is noted by C_k^{Sy}. The spectral density is the calculation of total signal power in a developed system. The spectral density formula is given in Eq. (26).

$$D = \frac{2ej}{o^2}\left(\left(Z_{real}(l)\right)^2 + \left(Z_{da}(l)\right)^2\right) \tag{26}$$

Here, the voltage signal is denoted by o^2.

4.2 Ceptral Feature Extraction

The gained speech signal S_{hd}^B is fed into the cepstral feature extirpation. The cepstral feature extirpation is a process of extirpating the best characteristics from noise signal information. The extracting features are MFCC and LPCC.

***MFCC*:** The MFCC is obtained from cepstral representation of speech signal. The MFCC is the collection of coefficients. The collections of audio frame coefficients are called MFCC. MFCC formula is calculated using Eq. (27).

$$z(o) = \left(0.54 - 0.46 * \cos\left(\frac{2\pi o}{O-1}\right) * t(o)\right) \tag{27}$$

Here, the value O is the signal's high level number of frames count. The convolution is noted by d. The usage energy is represented by I. The term M is the output and it is measured by Eq. (28).

$$M = \log_{10}\left(I * Y(:y) + 1e - 20\right) \tag{28}$$

Here, the variable e is the frequency. The variable Y is the signal. The extraction of MFCC provides greater accuracy during the speech recognition phase.

***LPCC*:** The LPCC is the collection of coefficients that calculate the current axis at a particular time period. The LPCC projected the previous voice signal's linear combinations also. The LPCC formula is calculated using Eq. (29).

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

$$L = L(o) + \sum_{l=1}^{o-1} \frac{l-o}{o} L(o-l)L(l) \tag{29}$$

Here, the variable z is the transform domain. The frames are noted by o. The coefficient of LPC is represented by l. These features are concatenated to get final features and it is noted by F_p^{Cp}.

4.3 Deep Feature Extraction

The gained speech signal S_{hd}^{B} is given to the autoencoder-related feature extirpation phase [28]. It is employed to extirpate the best variables of collected speech. The autoencoder contains two categories like, decoder and encoder phases. The encoder phase objective function is used to convert the high dimensional input into low dimensional with the help of the bottleneck layer. The high dimensional bottleneck features are transformed into low-dimensional features in the decoder layer. The function of the encoded operation is measured using Eq. (30).

$$z = g(\lambda, y) = t(Xy + c) \tag{30}$$

Here, the parameter z is the feature of vector. The parameter y is the input vector feature. The auto encoder's bias and weights are represented by c and X, respectively. The activation operation is represented by t. In the decoder phase, the hidden layer is mapped into the output of high dimensional features. The value a is the final outcome gate and it is given in Eq. (31).

$$a = h(\lambda`, z) = t(X`z + c`) \tag{31}$$

Here, the decoder and encoder outputs are noted by $\{\lambda, \lambda`\}$ and $a = h(\lambda` : (g(\lambda; y)))$, respectively. The target is denoted by y and the result of the network is represented by a. The extracted deep features from autoencoder are represented by A_e^{Dp}.

4.4 Optimal Weighted Feature Selection and Fusion

Feature Selection is a method remove the noise from the raw information. After dimensionality reduction, feature fusion assists in fully learning data features for the description of rich internal information. The extracted features from spectral C_k^{Sy}, cepstral F_p^{Cp} and deep features A_e^{Dp} are given into the optimal weighted feature selection phase. The suggested HBDOA algorithm is employed to optimally select the best parameters from spectral, cepstral and deep features. The variable LP_f^{Spect} is the selected features from spectral features, the variable FK_r^{Cept} is the chosen features from cepstral features and the variable SL_s^{Deep} is the selected features from deep features. The feature concatenation formula is given in Eq. (32).

$$Fk_1^t = TR_w^S * LP_f^{Spect} + \left(1 - TR_w^S\right) * FK_r^{Cept} \tag{32}$$

Here, the term Fk_1^t is the final feature-1. The term TR_w^S and SH_w^C are the optimized weights from developed HBDOA. The term FR_{fin}^T is calculated using Eq. (33).

$$FR_{fin}^T = SH_w^C * Fk_1^t + \left(1 - SH_w^C\right) * SL_s^{Deep} \tag{33}$$

Here, the term FR_{fin}^T is the final best weighted values. Also, optimized weights maximize the correlation coefficient. The correlation between the features of same class is measured. The objective functions are given in Eq. (34).

$$Ob_f = \underset{\left\{LP_f^{Spect}, FK_r^{Cept}, SL_s^{Deep}, TR_w^S, SH_w^C\right\}}{\arg\min} \left(\frac{1}{CRR}\right) \tag{34}$$

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

In feature selection, the variable LP_f^{Spect} is the selected features from spectral features and it is chosen in the range of $[1,50]$. The parameter FK_r^{Cept} is the selected features from cepstral features and it is getting in the interval of $[51,100]$. The variable SL_s^{Deep} is the selected features from deep features and it is getting in the range of $[101,150]$. The parameter TR_w^S and SH_w^C are the optimized weights and it is getting in the range of $[0.01,0.99]$. The formula of correlation coefficient is calculated by Eq. (35).

$$CRR = \frac{\Sigma\left(dk-\bar{d}\right)\left(ok-\bar{o}\right)}{\sqrt{\Sigma\left(dk-\bar{d}\right)^2\Sigma\left(ok-\bar{o}\right)^2}} \tag{35}$$

The variable $\bar{d}$ is the mean parameter. The terms dk and ok are the sample parameters. The correlation coefficient calculates the two parameter's statistical measures. The diagrammatical view of best weighted feature selection and fusion is displayed in Fig. 3.

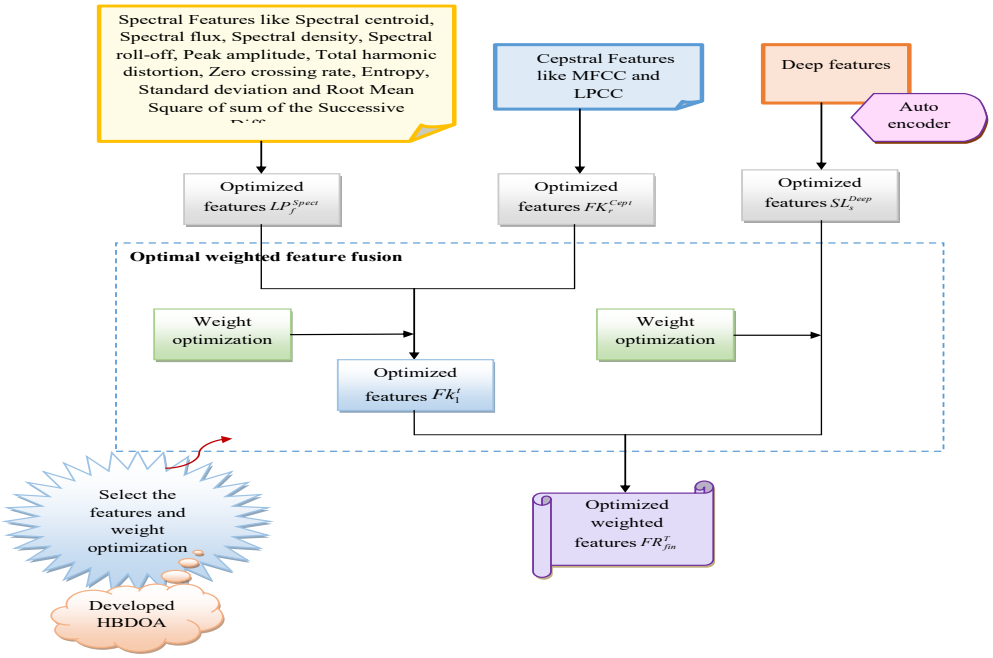


Figure 3. A diagrammatical view of best weighted feature selection and fusion

5 .Speech and Speaker Recognition using Attention-based Parameter Optimized Deep Learning Strategy

5.1 Basic BiLSTM

BiLSTM [29] contains four functions that is input gate, output gate, forget gate and weight measures. The BiLSTM is the effective techniques of RNN. It provided better outcomes for the detection and classification applications. It is the long-term implementation of time series data. The input is validated using Eq. (36).

$$k_v = \varphi\left(Y_{zk}z_v + Y_{jk}i_{v-1} + Y_{ek}E_{v-1} + d_k\right) \tag{36}$$

The output is measured using Eq. (37).

$$h_v = \varphi\left(Y_{zh}z_v + Y_{jh}i_{v-1} + Y_{eh}E_{v-1} + d_h\right) \tag{37}$$

The forget gate is calculated by Eq. (38).

$$i_v = \tanh\left(Y_{zi}z_v + Y_{ji}i_{v-1} + Y_{ei}E_{v-1} + d_i\right) \tag{38}$$

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

The weight measure of input is validated by Eq. (39).

$$E_v = \left(k_v * i_v + h_v E_{v-1}\right) \tag{39}$$

The weight measure of output is measured by Eq. (40).

$$q_v = \varphi\left(Y_{zq} z_v + Y_{jq} i_{v-1} + Y_{eq} E_{v-1} + d_q\right) \tag{40}$$

The weight measure of forgetting is calculated by Eq. (41).

$$j_v = q_v * \tanh\left(E_v\right) \tag{41}$$

Here, the term φ is the activation operation. The memory cell is noted by e. The gates of forgetting, input and output gate is noted by h, k and q, respectively. The present time node is used to get the output of the BiLSTM. The backward and forward operations are utilized to generate the outputs. The output is calculated using Eq. (42).

$$j_k = \overrightarrow{j_k} \oplus \overleftarrow{j_k} \tag{42}$$

Here, the terms $\overrightarrow{j_k}$ and $\overleftarrow{j_k}$ are the forward and backward operations. The additions of these operations are used to generate the final output. A basic diagram of BiLSTM is displayed in Fig. 4.

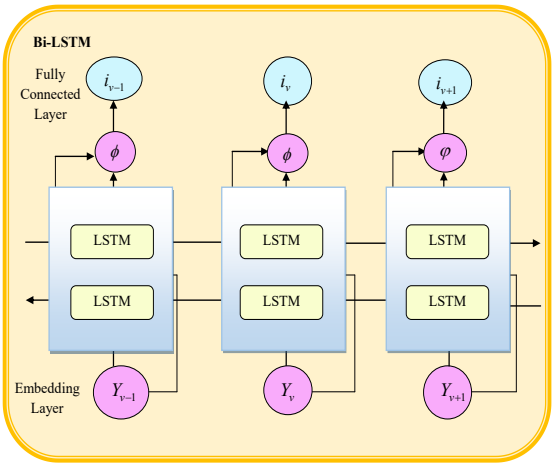


Figure 4 . Basic diagramtic representation of BiLSTM

5.2 TransBiLSTM-AM

The optimal weighted feature FR_{fin}^{T} is given to the TransBiLSTM-AM-based identification [30]. The transformer network contains two layers that are fully connected and point-wise layers. The decoder and encoder layer contains two sublayers. That is fully connected to feed-forward layer and multi-head-based self-attention. The transformer layer is validated by Eq. (43).

$$N = LayrNor\left(y + sublayer\left(y\right)\right) \tag{43}$$

Here, the variable $sublayer\left(y\right)$ is the sub layer. The variable $LayrNor$ is the normalized function. In the transformer, BiLSTM, the BiLSTM layers are added into the transformer encoder phase. Then, add the linear layer in the BiLSTM's decoder phase. Finally, the BiLSTM decoder phase helped to perform the prediction task. The BiLSTM output dimension size is set to 2.

***Attention*:** The attention method is an iterative process. The deep learning models are used to provide additional focus when using the attention mechanism. The attention mechanism reduces the information loss and implements the complex sequence also. The attention in BiLSTM is calculated using Eq. (44).

$$g_w = attent\left(y_w, u_{w-1}, \beta_{w-1}\right) \tag{44}$$

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Here, the parameter y_w is the input value and the parameter β_{w-1} is the weights of attention. The feed forward operation is represented by $attent()$. The frame of the previous operation is indicated by u_{w-1}. The attention weight is measured by Eq. (45).

$$\beta_{wn} = \frac{\exp(g_{wn})}{\sum_{n=1}^{N} \exp(g_{wn})} \tag{45}$$

The input is calculated using Eq. (46).

$$\hat{y}_w = \beta_{w-1} y_w \tag{46}$$

The output is measured by Eq. (47).

$$q\left(\frac{v}{y_w}\right) = BiLSTM(\hat{y}_w) \tag{47}$$

The term $BiLSTM(.)$ is the BiLSTM prediction scores and the attention scores are denoted by g_w. The term g_{wn} is the normalized attention scores.

5.3 Developed TransBiLSTM-AM for Speech and Speaker Recognition

The suggested TransBiLSTM-AM is employed to forecast the individual's voice for enhancing security systems. The suggested algorithm is employed to optimize the variables like number of epochs, activation function and hidden neuron count to enhance the accuracy and decrease the FPR, FDR, and FNR. The TransBiLSTM method decreased the exploding gradient issue. It reduces the computational time. But, it is expensive and it takes more time to implement the model in terms of the complex extracted features. Hence, the TransBiLSTM-AM method is implemented to resolve the recent issues of existing methods. Improving the accuracy and decreasing the FDR, FNR and FPR in the objective functions are calculated using Eq. (48).

$$Ob_f = \underset{\left\{JO_H^{Tr-BiLSTM}, BH_E^{Tr-BiLSTM}, VP_A^{Tr-BiLSTM}\right\}}{\arg\min} \left(\frac{1}{acy} + fnr + fpr + fdr\right) \tag{48}$$

In identification phase, the term $JO_H^{Tr-BiLSTM}$ is the optimized hidden neuron count and it is selected in the interval of $[5,255]$. The term $BH_E^{Tr-BiLSTM}$ is the optimized epoch count is and it is chosen in the range of $[5,50]$. The term $VP_A^{Tr-BiLSTM}$ is the optimized activation function and it is selected in the interval of $[0,4]$. The accuracy is the used to calculate the correct prediction of the designed system. The accuracy formula is given in Eq. (49).

$$acy = \frac{(YT_b + NO_h)}{(YT_b + NO_h + YT_c + NO_i)} \tag{49}$$

Here, the parameter true negative and true positive noted as YT_c and YT_b, respectively. The parameters false negative and positive indicated as NO_h and NO_i respectively. The FNR formula is given in Eq. (50).

$$fnr = \frac{NO_h}{NO_h + YT_b} \tag{50}$$

The FNR value is used to measure the false negative value divided by positive value. The FPR parameter is measured by Eq. (51).

$$fpr = \frac{YT_c}{YT_c + NO_h} \tag{51}$$

The number of negative predictions in the system is known as FPR. The number of false positive and true positive measures is known as FDR parameter. The FDR formula is given in Eq. (52).

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

$$fdr = \frac{NO_h}{YT_c + NO_i} \quad (52)$$

The diagrammatic view of developed TransBiLSTM-AM for speech and speaker recognition is shown in Fig. 5.

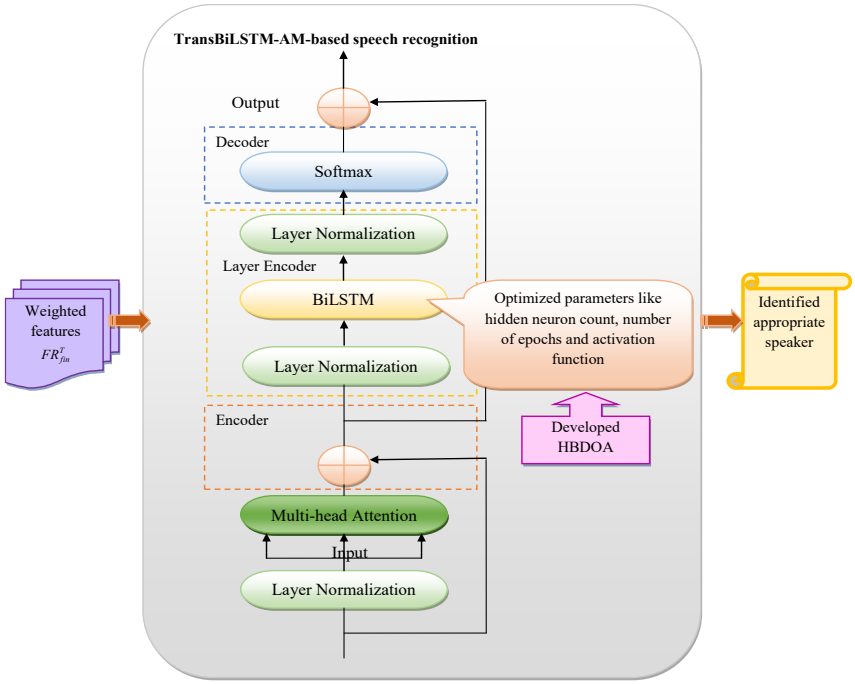


Figure 5. Structural demonstration of developed TransBiLSTM-AM for speech and speaker recognition

6. Results and Discussions

6.1 Experimental Setup

The investigated speech and speaker prediction system was developed using Python software. In terms of effectiveness measurements, the offered system's performance was compared with that of heuristic methodologies and conventional approaches. A chromosomal length of 3, a maximum iteration count of 50, a population size of 10, and were employed in the experimental investigation. The effectiveness of the developed models was compared using the existing approaches for the experimental study, including the Bidirectional Long Short-Term Memory (BiLSTM) [31], Transformer-BiLSTM (TransBiLSTM) [32] and Recurrent Neural Networks-LSTM (RNN-LSTM) [9]. The performance comparison also included the usage of heuristic algorithms like Moth-Flame Optimization (MFO) [33], Bat Algorithm (BAT) [34], BO [26], and DO [27].

6.2 Evaluation measures

The implemented speech and speaker recognition system employing some effectiveness metrics and is given in the below section.

(a) Accuracy: It is defined in Eq. (49).
(b) FNR: It is defined in Eq. (50).
(c) FPR: It is defined in Eq. (51).
(d) FDR: It is defined in Eq. (52).
(e) F1-score: $F1 = \frac{2 \times YT_b}{2YT_b + NO_h + NO_i}$

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

(f) Specificity: $SPE = \frac{NO_h}{NO_h + YT_c}$

(g) Sensitivity: $SEN = \frac{YT_b}{YT_b + NO_h}$

(h) Precision: $PRE = \frac{NO_h}{NO_i + YT_c}$

(i) NPV: $NPV = \frac{YT_c}{YT_c + NO_i}$

(j) FOR: $FOR = \frac{NO_i}{NO_i + YT_c}$

(k)MCC: $MCC = \frac{YT_b \times YT_c - NO_i \times NO_h}{\sqrt{(YT_b + NO_i)(YT_b + NO_h)(YT_c + NO_i)(YT_c + NO_h)}}$

(l) WER: The WER measures the correlation between word error rate and perplexity. The term WER is measured by Eq. (53).

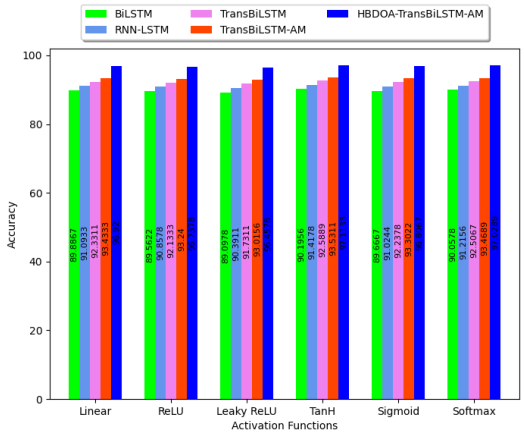
$$WER = \frac{T + E + J}{O} \tag{53}$$

Here, the parameter E is the total deletion count. The parameter T is the substitution. The correct words are noted by O. The total count insertions are indicated by J.

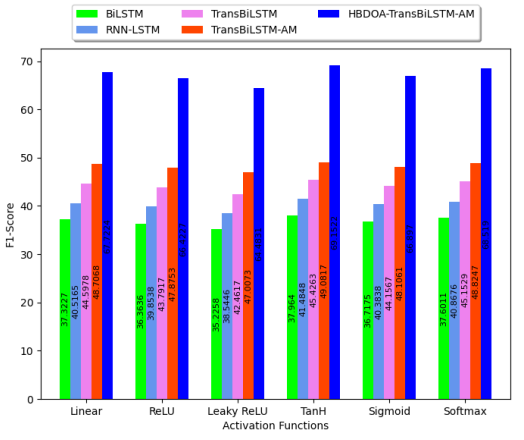
(l) Recognition rate: The microphone used to convert the vibration signal into electrical signal. This electrical signal accurate prediction is said to the recognition rate.

6.3 Performance analysis on the recommended speaker prediction system using dataset-1

The efficiency of the implemented speaker prediction model compared to traditional approaches and heuristic strategies is shown in Fig. 6 and 7, respectively. The HBDOA-TransBiLSTM-AM-based speaker recognition system showed a high accuracy of 12.03% than BiLSTM, 13.18% than RNN-LSTM, 11.42% than TransBiLSTM, and 26.77% than MVO-EDLN at tanh activation function. The suggested HBDOA-TransBiLSTM-AM-based speaker recognition system outperformed other models and showed great performance with regard to high accuracy.



(a)



(b)

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

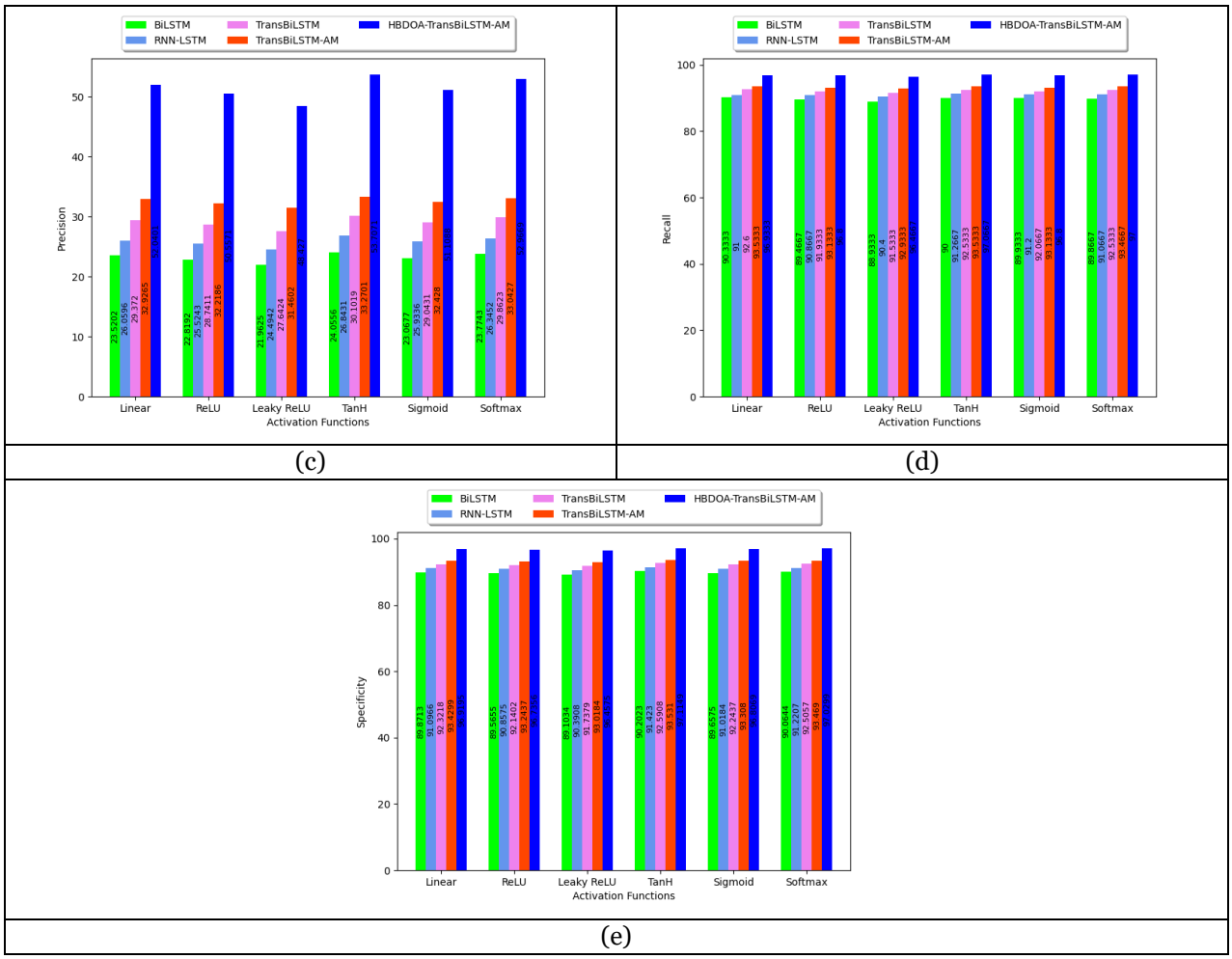
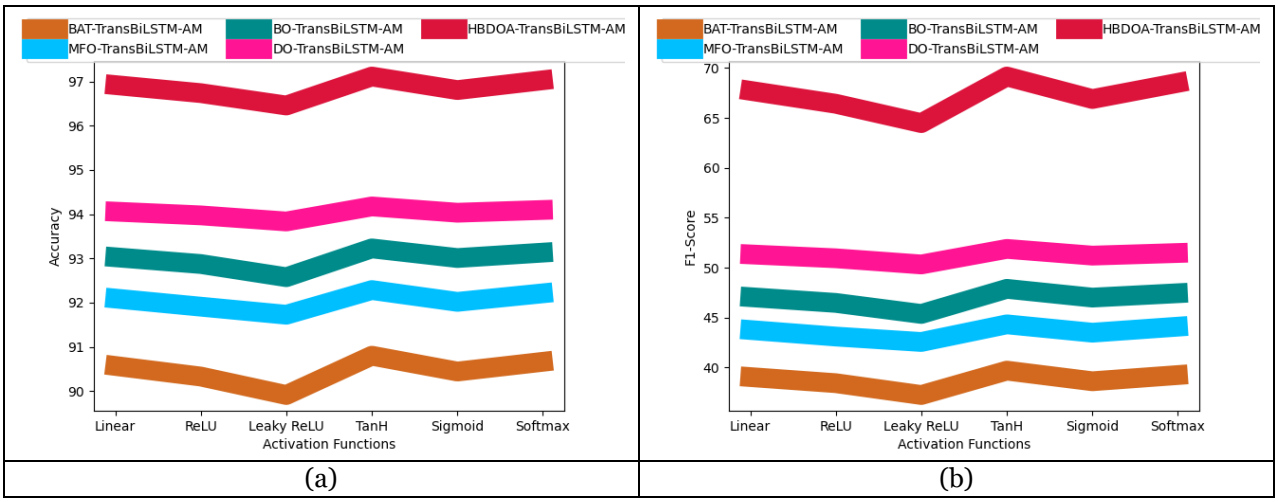


Figure 6 . Performance investigation on offered speaker recognition model utilizing deep learning over different methods with respect to (a) Accuracy (b) F1-score (c) Precision (d) Recall and (e) Specificity for dataset 1



Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

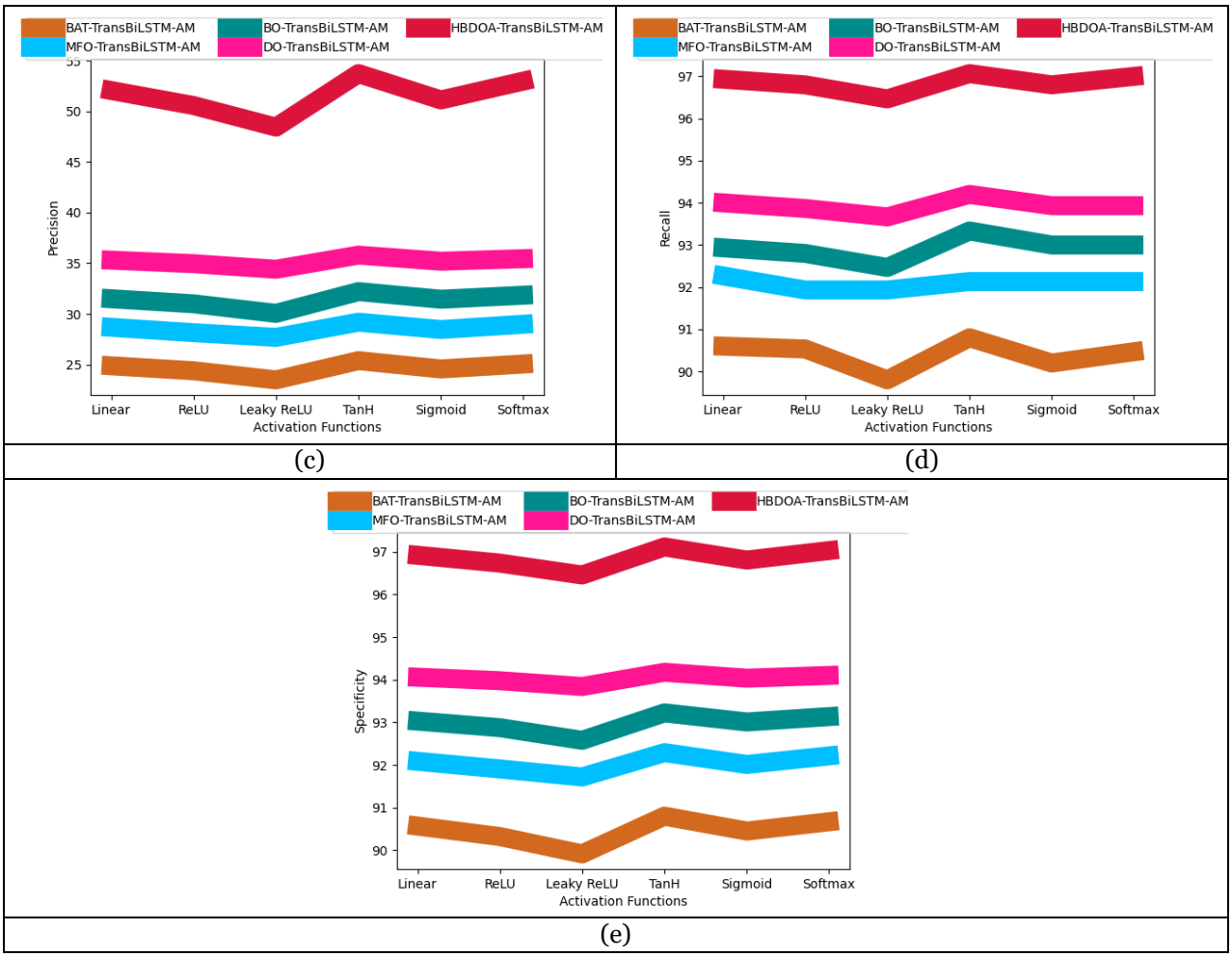


Figure 7 . Performance investigation on offered speaker recognition system utilizing deep learning over algorithms with respect to to (a) Accuracy (b) F1-score (c) Precision (d) Recall and (e) Specificity for dataset 1

6.4 Effectiveness testing on the recommended speaker prediction model using dataset-2

The suggested speaker recognition model efficiency compared to traditional approaches and heuristic algorithms is shown in Fig. 8 and 9, respectively. The HBDOA-TransBiLSTM-AM-based speaker recognition model provided a high f1-score of 16.93% than BiLSTM, 12.58% than RNN-LSTM, 10.42% than TransBiLSTM, and 50.66% than MVO-EDLN at activation function of ReLu. The suggested HBDOA-TransBiLSTM-AM-based speaker recognition system showed high performance with regard to high accuracy than other existing systems.

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

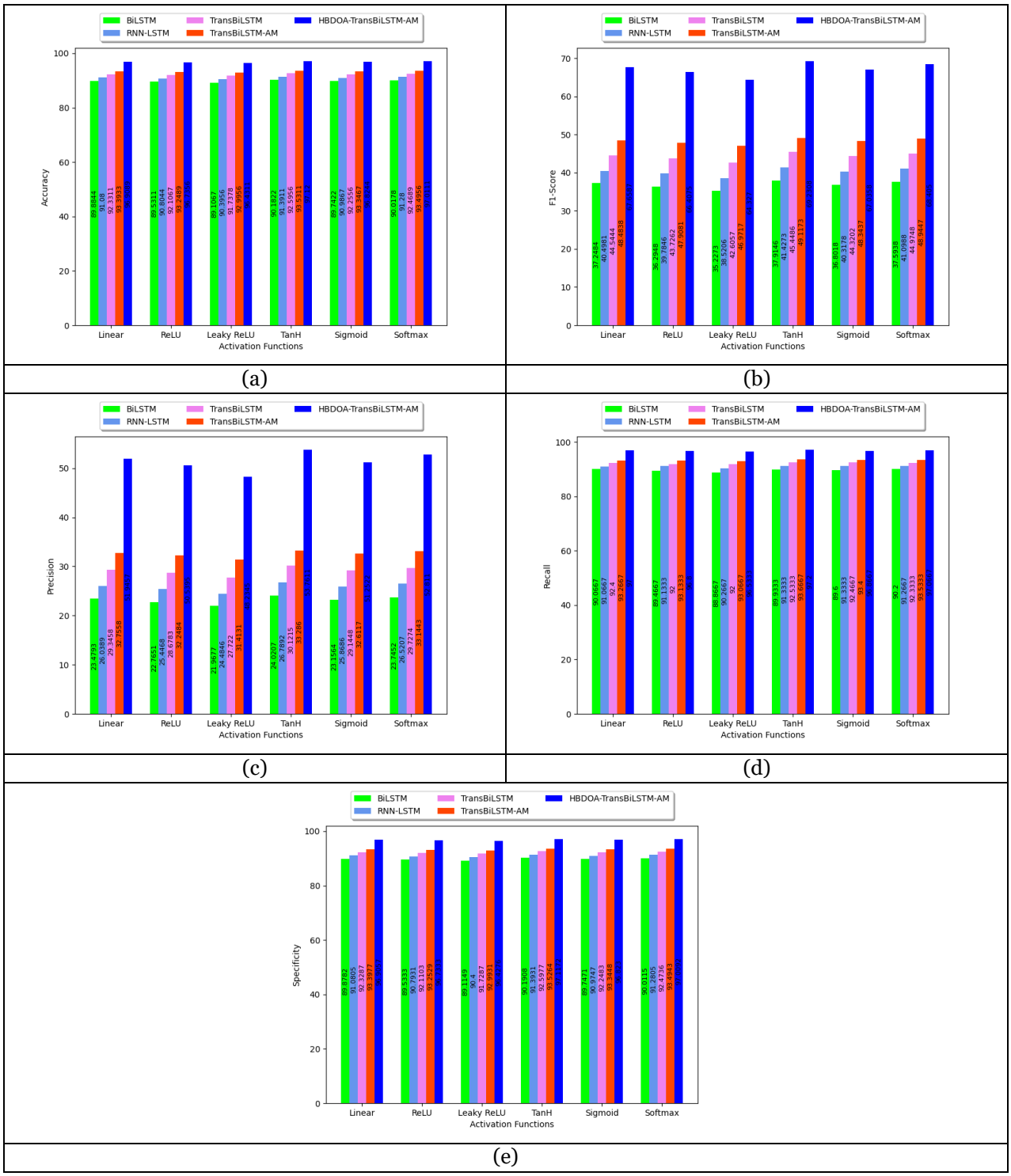


Figure 8 . Effectiveness investigation on offered speaker recognition model utilizing deep learning among methods with respect to (a) Accuracy (b) F1-score (c) Precision (d) Recall and (e) Specificity for dataset 2

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

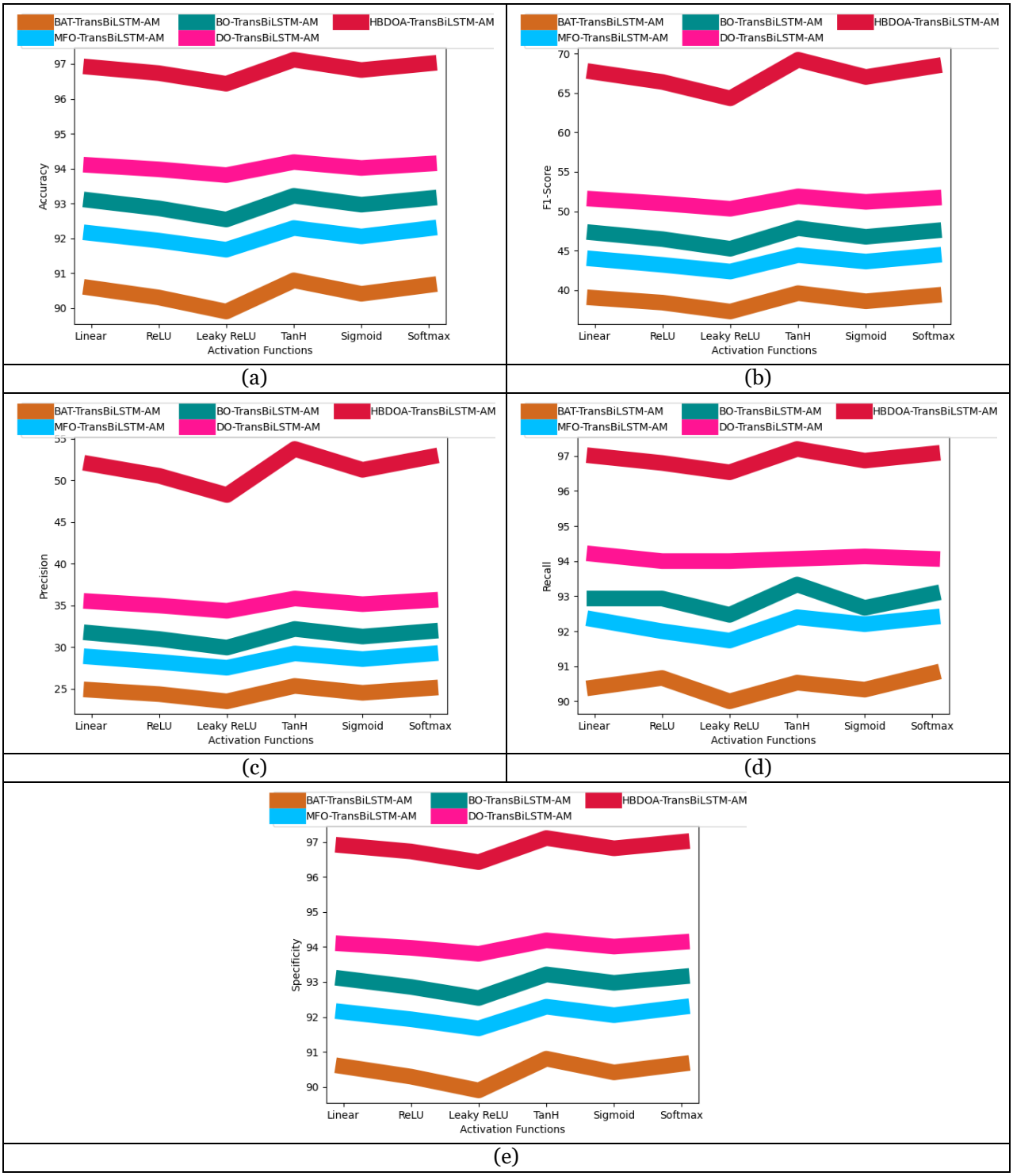


Figure 9 . Effectiveness investigation on offered speaker recognition model utilizing deep learning among strategies with respect to (a) Accuracy (b) F1-score (c) Precision (d) Recall and (e) Specificity for dataset 2

6.5 Performance analysis on the designed speech prediction model using dataset-1

Fig. 10 and Fig. 11, showed the efficiency of the implemented speech prediction model compared to heuristic strategies and traditional techniques. The HBDOA-TransBiLSTM-AM-related speech recognition model showed a high recognition rate of 18.43% than BiLSTM, 15.28% than RNN-LSTM,

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

10.52% than TransBiLSTM, and 16.77% than MVO-EDLN at activation function of tanh. The suggested HBDOA-TransBiLSTM-AM-based speech recognition system outperformed other models and provided great effectiveness with regard to high recognition rate.

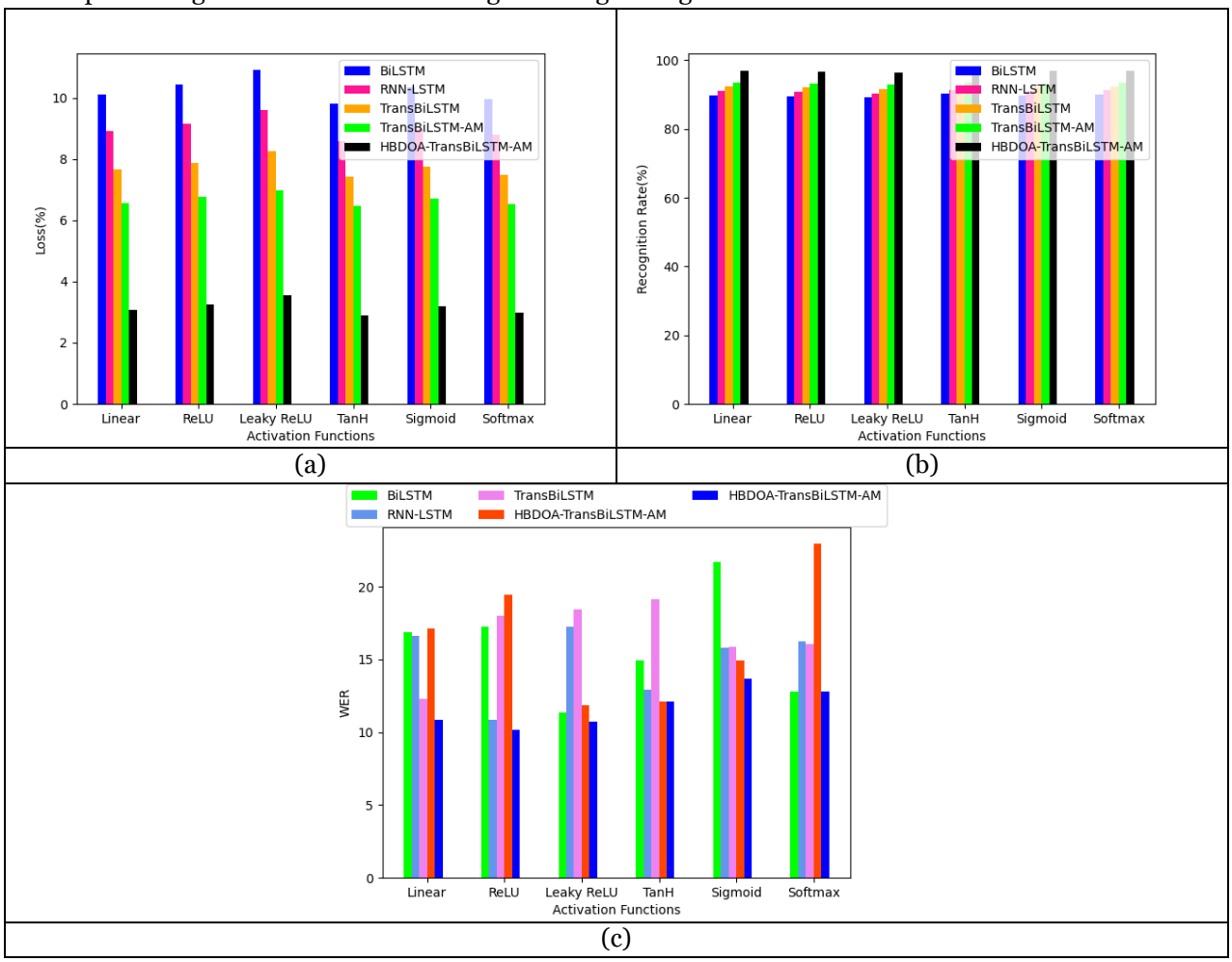
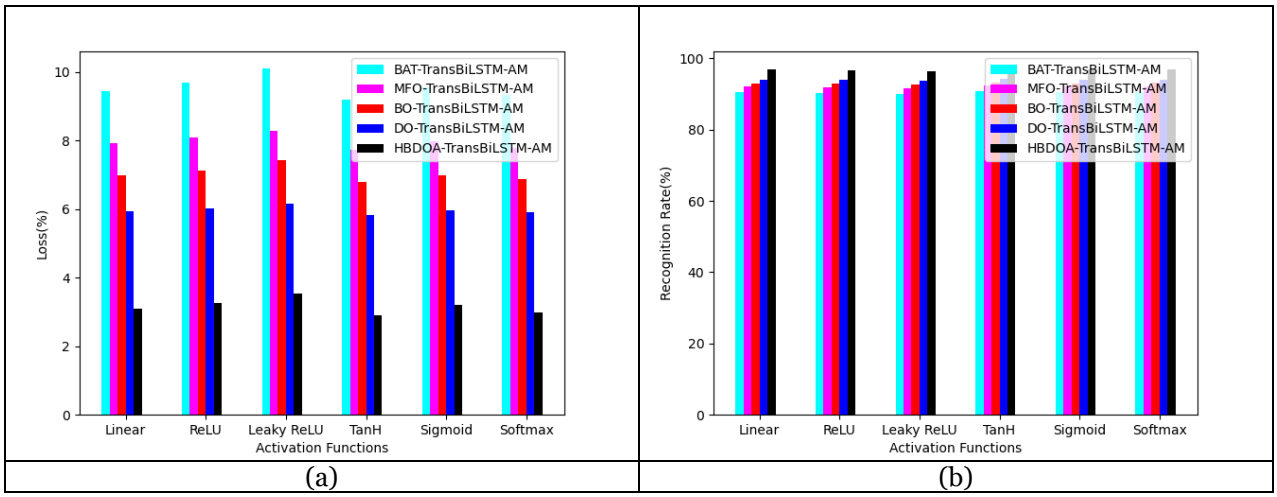
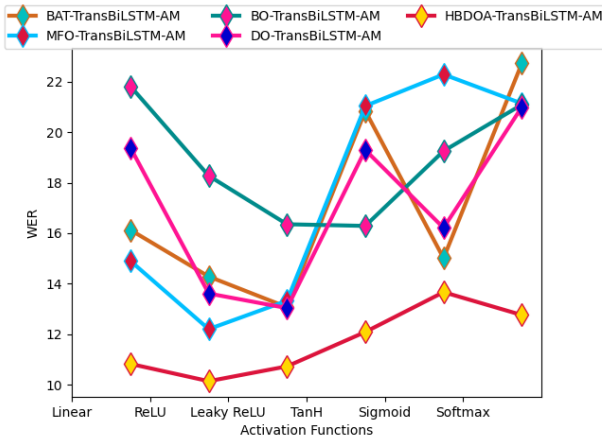


Figure 10. Performance investigation on offered speech recognition system utilizing deep learning over different methods with respect to (a) Loss (b) Recognition rate and (c) WER for dataset 1



Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

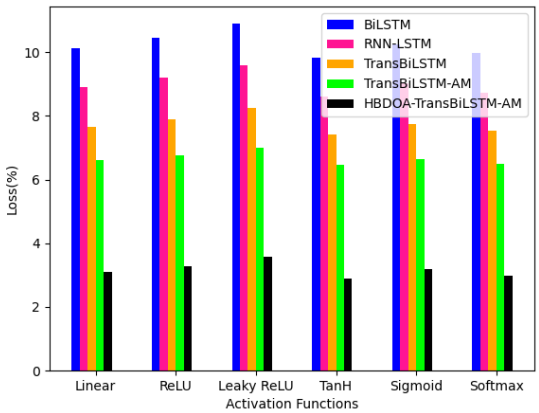


(e)

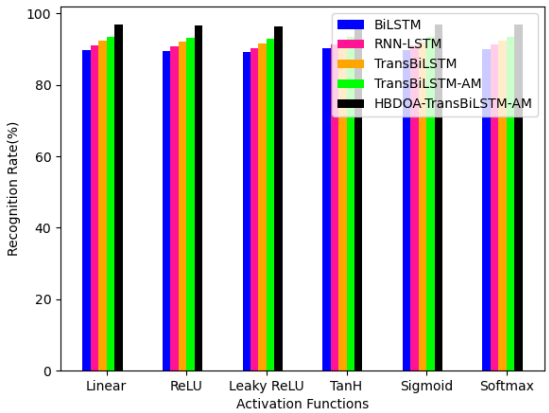
Figure 11. Performance investigation on offered speech recognition system utilizing deep learning over algorithms with respect to a) Loss (b) Recognition rate and (c) WER dataset 1

6.6 Performance investigation on the implemented speech prediction model using dataset-2

The suggested speech prediction model efficiency compared to traditional approaches and heuristic algorithms is shown in Fig. 12 and Fig 13, respectively. The HBDOA-TransBiLSTM-AM-based speech recognition system showed a high recognition rate of 14.13% than BiLSTM, 22.38% than RNN-LSTM, 9.42% than TransBiLSTM, and 40.66% than MVO-EDLN at activation function of ReLu. The suggested HBDOA-TransBiLSTM-AM-based speech recognition system showed high performance with regard to high recognition rate than other existing systems.



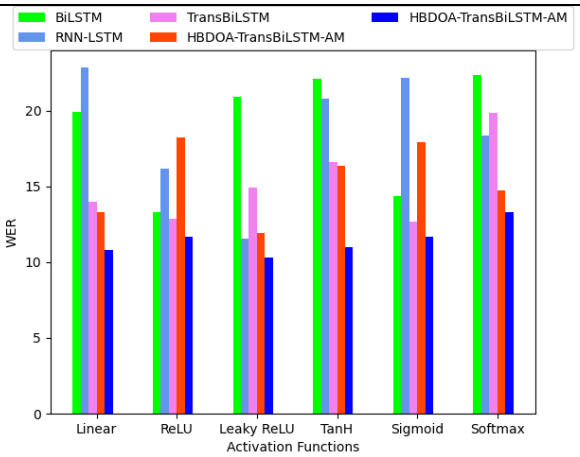
(a)



(b)

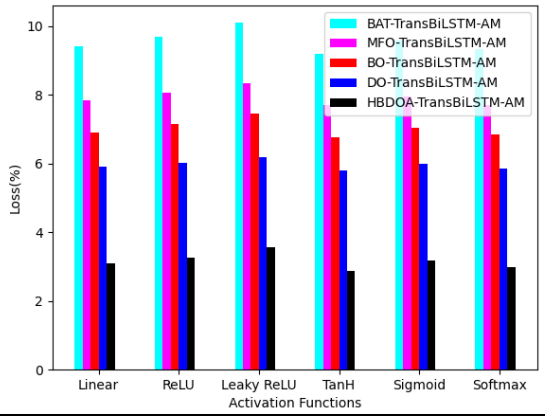
Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

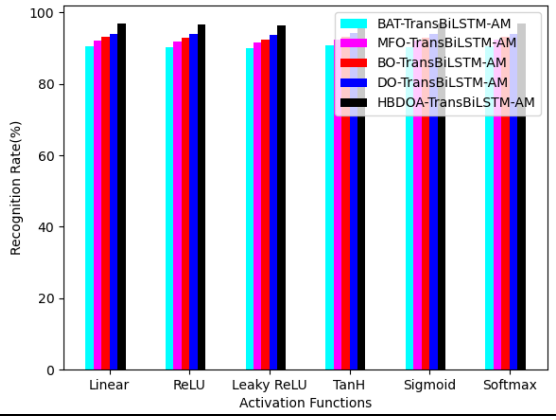


(c)

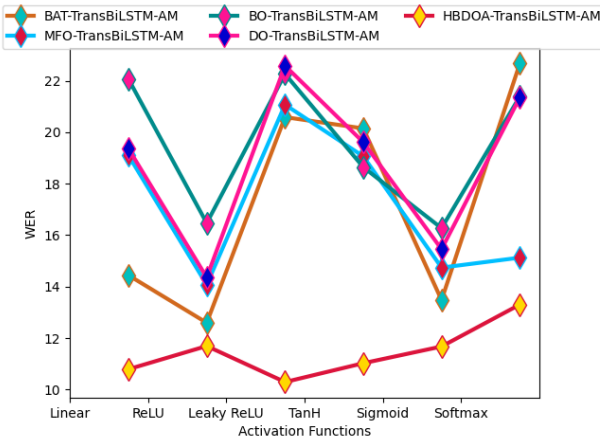
Figure 12 . Effectiveness investigation on offered speech recognition system utilizing deep learning over several methods with respect to (a) Loss (b) Recognition rate and (c) WER for dataset 2



(a)



(b)



(e)

Figure 13. Performance investigation on offered speech recognition system using deep learning over strategies with respect a) Loss (b) Recognition rate and (c) WER for dataset

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

6.7 Cost function investigation on the offered model

The investigated speech and speaker recognition system efficiency was compared to various heuristic strategies, as specified in Fig. 10. The HBDOA-TransBiLSTM-AM-based speech and speaker recognition system reduced cost function of 67.82% than BAT-TransBiLSTM-AM, 10.95% than MFO-TransBiLSTM-AM, 72.65% than BO-TransBiLSTM-AM, and 20.69% than DO-TransBiLSTM-AM using dataset-2 at the iterative parameter of 30. The HBDOA-TransBiLSTM-AM-related speech and speaker recognition system had given less cost function than existing systems for the experimental investigation.

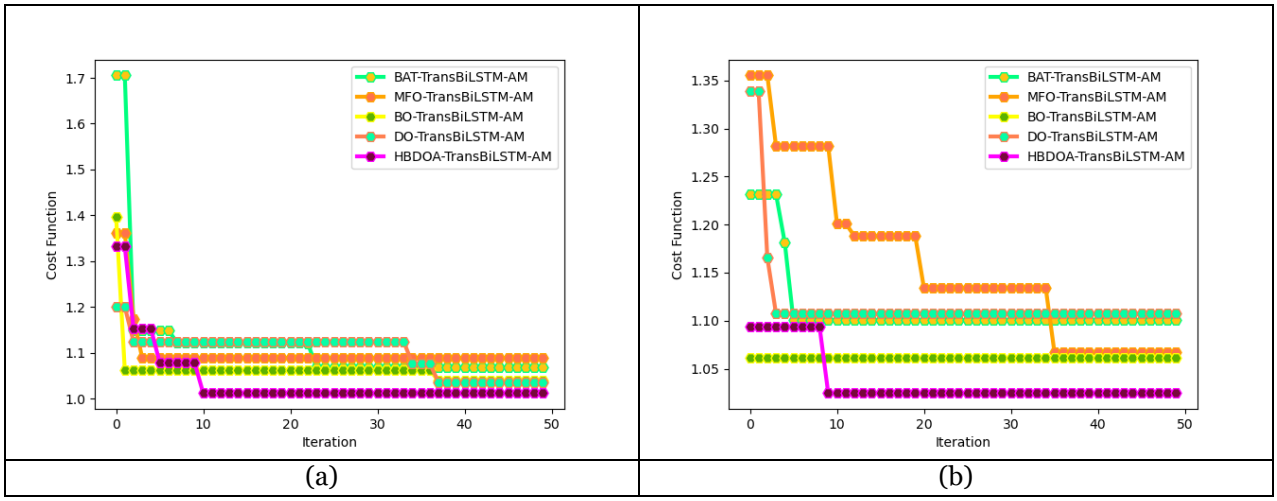


Figure 14 . Cost function testing on developed speech and speaker recognition system over several algorithms with respect to (a) dataset-1 and (b) dataset-2

6.8 ROC testing on the offered system

Fig. 11, showed the effective comparison of the investigated speech and speaker recognition system among various prediction techniques. The HBDOA-TransBiLSTM-AM-based speech and speaker recognition system provided high ROC of 18.13% than BiLSTM, 10.08% than RNN-LSTM, 28.62% than TransBiLSTM, and 16.22% than MVO-EDLN at the value of 0.4. The implemented speech and speaker recognition model provided better efficacy than conventional systems.

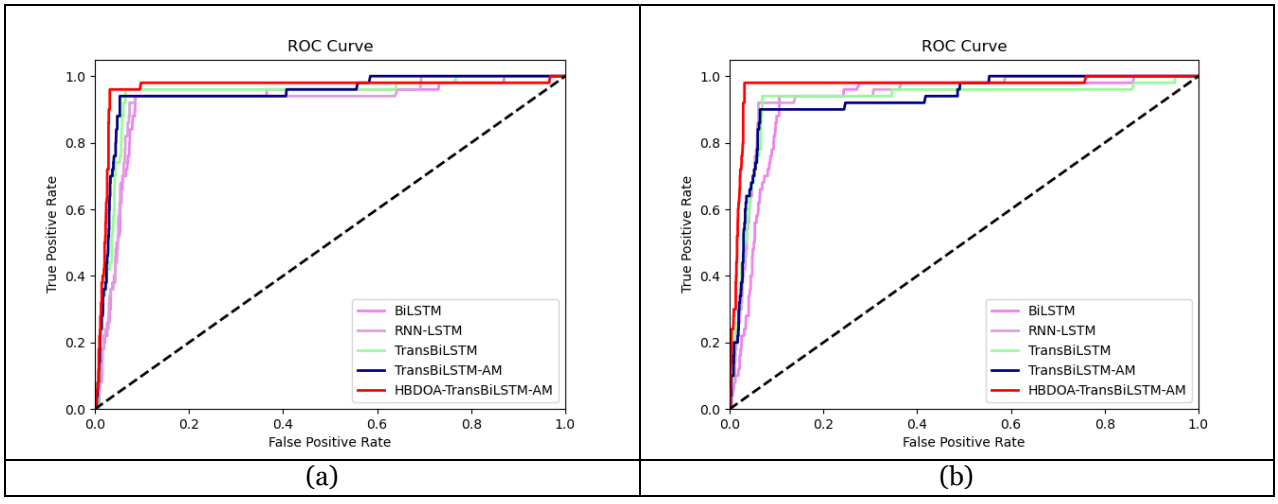


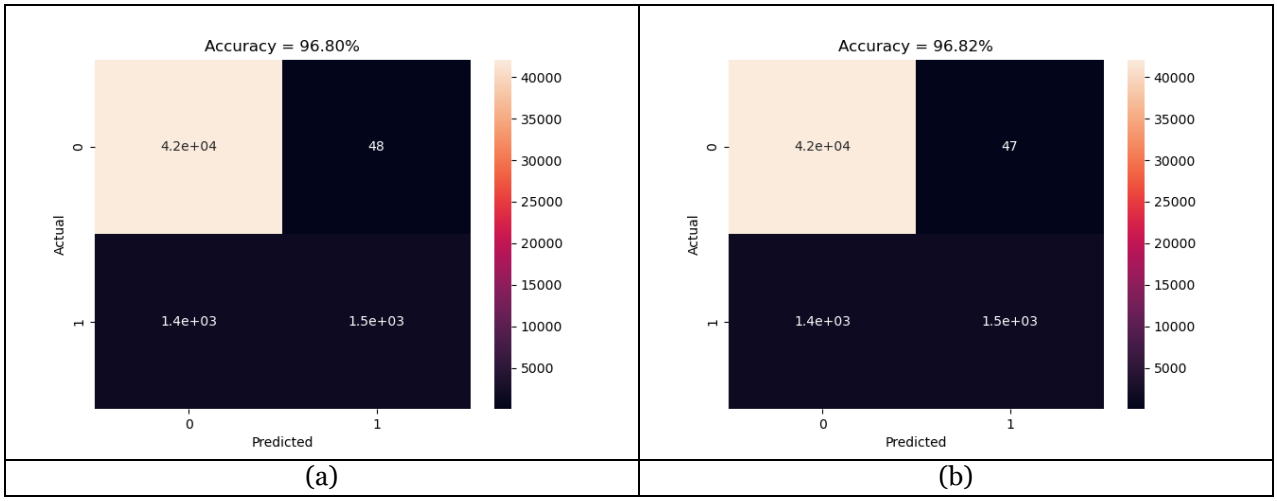
Figure 15. ROC testing on designed speech and speaker recognition system among several approaches with respect to (a) dataset-1 and (b) dataset-2

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

6.9 Confusion mattrix valuation on the suggested model

The confusion matrix analysis of speaker recognition model effectiveness was compared to various heuristic methods and techniques, which is shown in Fig. 12. The HBDOA-TransBiLSTM-AM-based speech and speaker recognition system performance showed high accuracy of 96.80%. The HBDOA-TransBiLSTM-AM-based speech and speaker recognition system outperformed previous prediction systems.



(a) (b)

Figure 16. Confusion mattrix of designed speech and speaker recognition model over several strategies and methods with respect to (a) dataset-1 and (b) dataset-2

6.10 Overall testing of the designed speech and speaker recognition system

The proposed speech and speaker recognition system efficiency was compared to several algorithms and methods and it is specified in Table II and Table III, respectively. The suggested HBDOA-TransBiLSTM-AM-based speech and speaker recognition system improved high precision of 20.42% than BAT-TransBiLSTM-AM, 39.03% than TransBiLSTM-AM, 50.05% than BO-TransBiLSTM-AM, and 56.09% than DO-TransBiLSTM-AM. The HBDOA-TransBiLSTM-AM-related speech and speaker recognition system is considered greater performance with high precision.

Table 2. Effectiveness testing of designed speech and speaker recognition system among different strategies

Terms	BAT-TransBiLSTM-AM [34]	MFO-TransBiLSTM-AM [33]	BO-TransBiLSTM-AM [26]	DO-TransBiLSTM-AM [27]	HBDOA-TransBiLSTM-AM
Dataset-1					
Accuracy	90.44222	92.01778	93.00889	94.03333	96.80667
Recall	90.2	92.13333	93	93.93333	96.8
Specificity	90.45057	92.01379	93.0092	94.03678	96.8069
Precision	24.56873	28.45964	31.44725	35.1986	51.10876
FPR	9.549425	7.986207	6.990805	5.963218	3.193103
FNR	9.8	7.866667	7	6.066667	3.2
FOR	0.372218	0.293942	0.25885	0.221967	0.113855
NPV	99.62778	99.70606	99.74115	99.77803	99.88615
FDR	75.43127	71.54036	68.55275	64.8014	48.89124
F1-Score	38.61852	43.48647	47.00135	51.20843	66.89703
MCC	0.441754	0.486833	0.517928	0.554698	0.690903
Dataset-2					
Accuracy	90.40889	92.05333	92.96444	94.01556	96.82444

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Recall	90.33333	92.2	92.66667	94.13333	96.86667
Specificity	90.41149	92.04828	92.97471	94.01149	96.82299
Precision	24.52045	28.56258	31.26406	35.15061	51.2522
FPR	9.588506	7.951724	7.025287	5.988506	3.177011
FNR	9.666667	7.8	7.333333	5.866667	3.133333
FOR	0.36733	0.291349	0.271243	0.214723	0.111467
NPV	99.63267	99.70865	99.72876	99.78528	99.88853
FDR	75.47955	71.43742	68.73594	64.84939	48.7478
F1-Score	38.57102	43.614	46.75412	51.18724	67.03576
MCC	0.441615	0.488037	0.515196	0.554925	0.692196

Table 3.performance estimation of suggested speech and speaker recognition model among different Methods

Terms	BiLSTM [31]	RNN-LSTM [9]	TransBiLSTM [32]	TransBiLSTM-AM [30]	HBDOA-TransBiLSTM-AM
Dataset-1					
Accuracy	89.66667	91.02444	92.23778	93.30222	96.80667
Recall	89.93333	91.2	92.06667	93.13333	96.8
Specificity	89.65747	91.01839	92.24368	93.30805	96.8069
Precision	23.06772	25.93365	29.04311	32.42804	51.10876
FPR	10.34253	8.981609	7.756322	6.691954	3.193103
FNR	10.06667	8.8	7.933333	6.866667	3.2
FOR	0.385676	0.332284	0.295689	0.253121	0.113855
NPV	99.61432	99.66772	99.70431	99.74688	99.88615
FDR	76.93228	74.06635	70.95689	67.57196	48.89124
F1-Score	36.71747	40.38376	44.15667	48.10606	66.89703
MCC	0.424886	0.458792	0.492311	0.527375	0.690903
Dataset-2					
Accuracy	89.74222	90.98667	92.25556	93.34667	96.82444
Recall	89.6	91.33333	92.46667	93.4	96.86667
Specificity	89.74713	90.97471	92.24828	93.34483	96.82299
Precision	23.15644	25.86858	29.14478	32.61173	51.2522
FPR	10.25287	9.025287	7.751724	6.655172	3.177011
FNR	10.4	8.666667	7.533333	6.6	3.133333
FOR	0.398	0.327423	0.280808	0.243219	0.111467
NPV	99.602	99.67258	99.71919	99.75678	99.88853
FDR	76.84356	74.13142	70.85522	67.38827	48.7478
F1-Score	36.80175	40.31783	44.32018	48.34369	67.03576
MCC	0.424949	0.458502	0.494491	0.529887	0.692196

6.11 Statistical estimation among heuristic algorithms

The effectiveness of the designed speech and speaker prediction system was compared over several heuristic strategies and it is depicted in Table IV. The suggested HBDOA-TransBiLSTM-AM-based speech and speaker recognition system improved best value of 30.12% than BAT-TransBiLSTM-AM, 10.73% than MFO-TransBiLSTM-AM, 12.15% than BO-TransBiLSTM-AM, and 21.44% than DO-TransBiLSTM-AM using dataset-2. The suggested speech and speaker recognition system outperformed than previous detection models.

Journal of Information Systems Engineering and Management
2026, 11(1s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Table 4.Statistical validation of offered speech and speaker recognition model over various strategies

Terms	BAT-TransBiLSTM-AM [34]	MFO-TransBiLSTM-AM [33]	BO-TransBiLSTM-AM [26]	DO-TransBiLSTM-AM [27]	HBDOA-TransBiLSTM-AM
Dataset-1					
Mean	0.930351	0.886214	0.884553	0.995778	0.965098
Worst	3.794138	2.938147	2.527495	5.985618	4.706301
Best	1.50529	1.27785	1.276206	1.320873	1.264794
Standard Deviation	1.281156	1.251954	1.212605	1.168947	1.1447
Median	0.440511	0.58853	0.513341	0.921169	0.575909
Dataset-2					
Mean	1.091086	0.893119	0.865986	1.08748	0.938416
Worst	6.200344	2.267646	3.845125	3.642113	5.540035
Best	1.391971	1.145949	1.243173	1.673581	1.536489
Standard Deviation	1.190396	1.273599	1.199844	1.08748	0.993035
Median	0.834698	0.294835	0.636286	1.004678	1.054541

7.Conclusion

The new suggested speech and speaker prediction system was used to identify the person through their voice with higher accuracy. The speech and speaker data was collected from the internet. The features were extracted from the collected data. The spectral removed features like standard deviation, spectral flux, spectral centroid, zero crossing rate, spectral density, peak amplitude, spectral roll-off, entropy, total harmonic distortion and RMSE sum of the succeeding were extracted. The cepstral features like MFCC and LPCC were removed. The deep variables were removed using the autoencoder approach. These removed features were fed into the fused weighted feature selection. The suggested HBDOA strategy was used to optimally select the best features from spectral, cepstral and deep features. Also, it optimized the weights to increase the correlation coefficient. The selected parameters were concatenated with optimized weights to get fused weighted features. Then, the weighted features were fed into the prediction section. The detection was detected employing TransBiLSTM-AM. Here, the suggested HBDOA algorithm was employed to optimize the parameters like the epoch count, hidden neuron count and activation function to improve the accuracy and decrease the FPR, FNR and FDR. Finally, it identifies and verifies the person through their voice effectively. The suggested HBDOA-TransBiLSTM-AM-based speech and speaker recognition system improved high precision of 12.85% than BAT-TransBiLSTM-AM, 25.03% than TransBiLSTM-AM, 65.05% than BO-TransBiLSTM-AM, and 16.29% than DO-TransBiLSTM-AM. The suggested speech and speaker recognition system performance was compared among previously used models and it attained high accuracy.

Reference

[1] K. Azizah and W. Jatmiko, "Transfer Learning, Style Control, and Speaker Reconstruction Loss for Zero-Shot Multilingual Multi-Speaker Text-to-Speech on Low-Resource Languages," IEEE Access, vol. 10, pp. 5895-5911, 2022.

[2] O.Ghahabi and J. Hernando, "Deep Learning Backend for Single and Multisession i-Vector Speaker Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 4, pp. 807-817, April 2017.

[3] ShivaniGoel and Youddha Beer Singh, "An efficient algorithm for recognition of emotions from speaker and language independent speech using deep learning," Multimedia Tools and Applications, vol.80, pp.14001–14018,2021.

[4] Chin-HuiLeec,Sabato Marco Siniscalchi and ZhenHuang, "A unified approach to transfer learning of deep neural networks with applications to speaker adaptation in automatic speech recognition,"Neurocomputing,vol.218,pp.448-459, 19 December 2016.

[5] K. Chen and A. Salman," Learning Speaker-Specific Characteristics With a Deep Neural Architecture," IEEE Transactions on Neural Networks, vol.22, no.11, pp. 1744-1756, Nov 2011.

[6] O. Abdel-Hamid,L.Dai,H. Jiang,Q.Liu and S.Xue,"Fast Adaptation of Deep Neural Network Based on Discriminant Codes for Speech Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 12, pp. 1713-1725, Dec. 2014.

[7] M.Kim,Y.Kim, H.Kim, J.Wang and J.Yoo,"Regularized Speaker Adaptation of KL-HMM for Dysarthric Speech Recognition," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 25, no. 9, pp. 1581-1591, Sept. 2017.

[8] H. Meng, Wei and T. Yan,"Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network," IEEE Access, vol. 7, pp. 125868-125881, 2019.

[9] T.W. Sun,"End-to-End Speech Emotion Recognition with Gender Information," IEEE Access, vol. 8, pp. 152423-152438, 2020.

[10] ShivaniGoel and Youddha Beer Singh,"An efficient algorithm for recognition of emotions from speaker and language independent speech using deep learning,"Multimedia Tools and Applications,vol.80,pp.14001–14018,2021.

[11] G.Hiroshi, Kazuhiro Nakadai,KuniakiNoda,Okuno,Tetsuya Ogata and Yuki Yamaguchi," Audio-visual speech recognition using deep learning,"Applied Intelligence, vol.42, pp.722–737 ,2015.

[12] Gaurav Agarwal,Hari Om,Performance of deer hunting optimization based deep learning algorithm for speech emotion recognition,"Multimedia Tools and Applications vol.80, pp.9961–9992,2021.

[13] AkshayDeepak,AshishRanjan,GayadharPradhan,Md Shah Fahad,"Speaker Adversarial Neural Network (SANN) for Speaker-independent Speech Emotion Recognition ," Circuits, Systems and Signal Processing,vol. 41, pp.6113–6135, 2022.

[14] AswinShanmugam Subramanian,Chao Weng,Dong Yu,Meng Yu,Shinji Watanabe, "Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition,"Computer Speech and Language,vol.79, pp:101360, September 2022.

[15] A. A. Joshy and R. Rajan, "Automated Dysarthria Severity Classification: A Study on Acoustic Features and Deep Learning Techniques,"IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 30, pp. 1147-1157, 2022.

[16] V. Andrei, C. Burileanu and H. Cucu and, "Overlapped Speech Detection and Competing Speaker Counting–-Humans Versus Deep Learning," IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 4, pp. 850-862, Aug 2019 .

[17] A. Amjad, N. Ashraf,H.T. Chang L. Khan, and M. B. Mahmood , "Recognizing Semi-Natural and Spontaneous Speech Emotions Using Deep Neural Networks," IEEE Access,vol. 10, pp. 37149-37163, 2022.

[18] Yesim Dokuz and Zekeriya Tufekci,"Mini-batch sample selection strategies for deep learning based speech recognition,"Applied Acoustics,vol.171, pp.107573,2021.

[19] Chin-Hui Lee,Feng Ma,Hai-Kun Wang,Jing-Dong Chen,Jun Du,Le Sun and Yan-Hui Tu,"An iterative mask estimation approach to deep learning based multi-channel speech recognition,"Speech Communication ,vol.106, pp.31-43, January 2019.

[20] Amod Kumar,Gurpreet Kaur and Mohit Srivastava,"Speaker and Speech

[21] Recognition using Deep Neural Network,"Article,June 2018.

[22] Seyed Reza Shahamir,"Speech Vision: An End-to-End Deep Learning-Based Dysarthric Automatic Speech Recognition System,"IEEE Transactions on Neural Systems and Rehabilitation Engineering,vol.29, 2021.

[23] AnanyaMisra,Andrew,Arun Narayanan,Bo Li,Chanwoo Kim,Ehsan Variani,IzhakShafran,Kean Chin,Kevin W.Wilson,MichielBacchiani,Ron J Weiss and Tara N. Sainath,"Multichannel Signal Processing With Deep Neural Networks for Automatic Speech Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 5, pp. 965-979, May 2017.

[24] Jilenkumari Devi,KhelchandraThongam and NangbamHerojit Singh,"Automatic Speaker Recognition from Speech Signals Using Self Organizing Feature Map and Hybrid Neural Network,"Microprocessors and Microsystems,vol.79, pp.103264, November 2020.

[25] Jamal Kharroubi&SoufianeHourri, "A deep learning approach for speaker recognition," International Journal of Speech Technology, vol. 23, pp.123–131, 2020.

[26] Ziye Yang, Shanzheng Guan and Xiao-LeiZhang, "Deep ad-hoc beamforming based on speaker extraction for target-dependent speech separation," Speech Communication, vol. 140, pp. 87-97, May 2022.

[27] Amit Kumar Das and Dilip Kumar Pratihar, "A New Bonobo Optimizer (BO) for Real-Parameter Optimization," Conference: The IEEE Region 10 Symposium, TENSYMP 2019.

[28] Xiguang Li, Shoufei Han, Liang Zhao, Changqing Gong, and Xiaojing Liu, "New Dandelion Algorithm Optimizes Extreme Learning Machine for Biomedical Classification Problems," Computational Intelligence and Neuroscience, Article ID 4523754, pp.13, 2017.

[29] Bhavik Vachhani,Chitralekha Bhat,Biswajit Das, INRIA Bordeaux, Sunil Kumar Kopparapu,"Deep Autoencoder Based Speech Features for Improved Dysarthric Speech Recognition", Stockholm, pp.20-24, August 2017.

[30] Dongyu Du, Yichen Ji, Lianjie Li, "BELSTM: Understanding the Transformer and Bidirectional Long Short-Term Memory for Early Rumor Detection" , Electronic Information Technology and Computer Engineering pp. 687–691, November 2020.

[31] Zhiheng Huang, Peng Xu, Davis Liang, Ajay Mishra, Bing Xiang, "TRANS-BLSTM: Transformer with Bidirectional LSTM for Language Understanding", Computation and Language, pp. 20, 16 march 2020.

[32] D. Yoon, Z. Yeeh and J. Byun, "Seismic Data Reconstruction Using Deep Bidirectional Long Short-Term Memory With Skip Connections," IEEE Geoscience and Remote Sensing Letters, vol. 18, no. 7, pp. 1298-1302, July 2021.

[33] Z. Liu, X. Kang and F. Ren, "Dual-TBNet: Improving the Robustness of Speech Features via Dual-Transformer-BiLSTM for Speech Emotion Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2193-2203, 2023.

[34] G. Chandrasekaran, N. S. Kumar, P. R. Karthikeyan, K. Vanchinathan, N. Priyadarshi and B. Twala, "Test Scheduling and Test Time Minimization of System-on-Chip Using Modified BAT Algorithm," IEEE Access, vol. 10, pp. 126199-126216, 2022.

[35] G. Chandrasekaran, N. S. Kumar, P. R. Karthikeyan, K. Vanchinathan, N. Priyadarshi and B. Twala, "Test Scheduling and Test Time Minimization of System-on-Chip Using Modified BAT Algorithm," IEEE Access, vol. 10, pp. 126199-126216, 2022.