International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

An Automated DTFWOA-ASI model for Classification and Identification of Speakers utilizing a Metaheuristic Algorithm

T.S. Mullai vendan*[1], R. Thiruvengatanadhan[2], P. Dhanalakshmi[3]

Submitted: 27/01/2024 **Revised**: 05/03/2024 **Accepted**: 13/03/2024

Abstract: Identifying a speaker is a task of classification that aims at recognizing an individual based on sequential data over time. Given that a speech signal manifests as a one-dimensional, continuous time series, most contemporary studies rely on either convolutional neural networks (CNN) or recurrent neural networks (RNN). These techniques have shown effectiveness across various applications, yet efforts to merge these two models for investigating speaker recognition tasks remain unexplored. A spectrogram integrated into a speech signal reveals the voiceprint's spatial attributes, reflecting the voice spectrum. This makes CNN highly suitable for drawing out spatial characteristics, essentially capturing the spectral correlations present in acoustic signatures. Concurrently, with the speech signal being time-sequential, deep RNNs are superior in depicting extended speech compared to more superficial networks. The study introduces a new model named Dual-Tier Feature Extraction with Whale Optimization Algorithm for Automated Speaker Identification (DTFWOA-ASI), designed to address the shortcomings found in earlier models. The DTFWOA-ASI approach is a cutting-edge method designed specifically for the identification of speaker identities. It employs the method of average median filtering (AMF) to remove background noise from sound recordings. Subsequently, the strategy utilizes both MFCC and spectrogram data as inputs into the VGGish model, an advanced deep-learning convolutional network engineered for extracting crucial features. For the fine-tuning of the LSTM-RNN model's hyperparameters, the technique makes use of the Whale Optimization Algorithm (WOA). The approach integrates a long short-term memory network with a recurrent neural network (LSTM-RNN) to enable the automatic identification and classification of speakers. The performance and accuracy of the DTFWOA-ASI framework were thoroughly assessed through several experimental procedures. A comparative analysis highlights the model's superior performance in comparison to the latest methodologies.

Keywords: *Speaker identification, Deep Learning, Whale optimization algorithm, VGGish, Spectrograms.*

1. Introduction

Identifying a person through their voice is what speaker recognition is all about. The uniqueness of every voice comes from variables such as the larynx size, How the vocal tract is shaped, and various components involved in creating voice sounds [1]. This method poses challenges because individuals being tested do not reveal their identities, necessitating a comparison between one and N, wherein N symbolizes the aggregate count of people who have been enrolled. In contrast, speaker verification concerns itself with confirming if a person is indeed who they profess to be. Given that any false identities are considered unknown, this procedure is often referred to as open-set recognition [2]. Conversely, when it's determined that a voice belongs to a registered individual, this scenario is termed closed-set recognition. Speech carries a wealth of information, including emotional nuances, accent, and gender among other traits, making it a potent medium for communication [3]. These distinguishing characteristics allow for the pinpointing of individuals through their voiceprints. The deep learning networks are then trained using these unique vocal samples. In the identification phase, the voice's features are compared against a database of models [4]. The speaker who most likely made the utterance is then pinpointed as the intended target [5].

Techniques for recognizing speakers automatically are proficient in determining an individual's identity through their vocal signals. Recognition technologies are widely used for a variety of purposes, including but not limited to - Securely granting access to different services such as voice mail and telephone banking, Facilitating voice-activated dialing for data networks, accessing databases and computers remotely, making use of information and shopping services via phone, enhancing security for web services and confidential data sectors, aiding law enforcement in surveillance activities, crime investigation, logging activities from remote locations, and monitoring phone calls within prisons [6]. For these systems to verify a speaker accurately, it's crucial to extract significant features from each speech segment, focusing on pivotal aspects of the vocal signal [7]. The MFCC, in particular, is widely adopted for its effectiveness in clean environments, marking a high rate of efficiency [8]. Nonetheless, the performance of the MFCC technique diminishes significantly in environments plagued by echo and background noise [9]. A challenge with contemporary

[1] Research Scholar, Department of Computer and Information Science, Annamalai University.
[2] Assistant Professor, Department of Computer Science and Engineering, Annamalai University.
[3]Professor, Department of Computer Science and Engineering, Annamalai University.
**Email:tsmullai24@gmail.com[1],thiruvengatanadhan01@gmail.com[2], abidhana01@gmail.com[3]*

methods is the necessity for input samples to be large enough for effective training and speaker recognition, which, in turn, hikes up the demand for computing resources [10]. Moreover, these methods struggle to deliver their best in noisy settings. Recent research presented a robust method for speaker identification, employing a hybrid approach to recognize speakers from their speech with high efficiency, despite background noise and other challenges [11].

This paper unveils a novel approach for Automated Speaker Identification, named Dual-Tier Feature Extraction enhanced by Whale Optimization Algorithm (DTFWOA-ASI). At the heart of the DTFWOA-ASI framework is the strategic use of the mean median filtering method, which effectively removes noise from audio signals. Once the audio has been cleared of disturbances, the approach proceeds by feeding both MFCC and spectrogram data into the VGGish model. This model is grounded in a deep convolutional neural network, serving as the cornerstone for extracting key features. Furthermore, the Whale Optimization (WOA) algorithm is instrumental in fine-tuning the hyperparameters of the VGGish model to enhance its performance. For precise automatic speech recognition classification, the framework cleverly combines a long short-term memory (LSTM) function with a recurrent neural network (RNN), ensuring a dynamic and powerful analysis tool. The performance and reliability of the DTFWOA-ASI model have been rigorously verified across a variety of experimental setups, demonstrating its effectiveness in achieving its objectives.

2. Related Works

In [12], the author scrutinized the effectiveness of incorporating the MFCC characteristic, derived either directly from the DWT framework or through feature warping, in enhancing the precision of speaker verification systems based on identity vectors (a-vectors), especially under conditions laden with reverberation and noise. This stratagem has been applied to heighten the performance of forensic speaker verification (FSR) and for the assembly of legal proofs in judicial settings. In a subsequent research endeavor, [13] unveiled an innovative approach known as MLHF-SVM, a tri-layer hybrid methodology incorporating fuzzy logic with support vector machine (SVM) tactics. This model is intricately designed across three distinct levels: The feature extraction phase, the Pre-classification stage, and the Classification level. The revolutionary MLHF-SVM approach addresses the previously highlighted challenges by integrating FCM-driven authentication details with a tiered arrangement of SVM classifiers. To surmount the FCM system's tendency towards being ensnared in local minimums, an enhanced strategy combining a novel approach of natural exponential inertia weight particle swarm optimization (IEPSO) with FCM has been initiated for optimal results.

In [14] the author presents a method employing a deep neural network (DNN) that merges the capabilities of a two-dimensional convolutional neural network (2D-CNN) and a gated recurrent unit (GRU) with the aim of speaker identification. This network's design utilizes a convolutional layer for reducing the size across both frequency and time dimensions, simultaneously pulling out characteristics from the voiceprints to facilitate quick processing by the GRU layer. Additionally, the GRU layers, which are stacked in a recurrent network, are adept at acquiring a speaker's unique acoustic characteristics. In another study, the author [15] was able to distinguish a phone call recording from various unforeseen sound clips by employing a support vector machine (SVM) model. Saleem and his collaborators [16] unveiled an innovative feature set reduction (FSR) approach that relies on gleaning language and accent information from brief spoken segments. To automate these tasks, they implemented a variety of standard and deep learning (DL) techniques. Among the new CNN-based frameworks implemented were the GMM-CNN and VGGVox, both of which made use of speech spectrograms for processing. When discussing DNN applications, an x-vector model was chosen, relying on DNN embeddings for its operations. The criticality of choosing the right filter features has been emphasized in [17]. It was found that combining filter feature selection with methods such as logistic regression, random forests, and KNN allows for the identification of essential acoustic characteristics. Also, research in [18] demonstrates that the inclusion of artificial noise and reverberation into training data significantly boosts the performance of DNN embedding systems. The study was based on the Cantonese segment of the NIST SRE 2016 evaluation (SRE16) and utilized speakers from natural settings. This inspired our team to conduct speaker verification experiments utilizing a database spontaneously generated under unregulated conditions.

In [20], the author introduced an innovative hybrid approach for ASR, leveraging an ANN model based on speech signals. They effectively made use of the Mel-frequency cepstral coefficients (MFCC) to extract features. These extracted features are then utilized as the input sample, which the Self-Organizing Feature Map (SOFM) processes further to diminish its dimensions. Ultimately, the MLP model equipped with Bayesian regularization facilitates the recognition process with this dimensionally reduced input. In a related development, the researcher identified as [21] has shown that it's possible to pull speaker embeddings in a manner that not only keeps the speaker's identity intact but also upholds the secrecy of the model owned by the service provider, all this made achievable through Secure Multiparty Computation.

Additionally, we present the possibility of achieving a reasonable balance between computational and security expenses. Our research complements existing studies by demonstrating the private implementation of verification, marking progress towards completely private ASR systems. Unlike previous studies that were conducted under controlled conditions based on certain premises, our experiments utilized a database compiled in an uncontrolled setting, yet we were able to maintain commendable levels of accuracy.

3. The Proposed DTFWOA-ASI model

In this paper, we introduce the new DTFWOA-ASI framework, focusing on speaker identification. This innovative model incorporates a combined development set comprising 7500 instances collected from five different speakers, with the audio recorded at a 16 kHz sampling rate. The dataset is split into training and validation parts, with 70% dedicated to training and the remaining 30% for validation purposes. Speech segments within this dataset vary, lasting from 3 to 5 seconds each. Initially, the DTFWOA-ASI framework employs the AMF method to eliminate noise from the audio signals. Subsequently, the VGGish model leveraged both the MFCC and spectrograms as input. The Whale Optimization Algorithm (WOA) is then leveraged for fine-tuning the hyperparameters within the LSTM-RNN model. In the final phase, the LSTM-RNN model is used as a tool for classifying automatic speech recognition (ASR) tasks. Fig 1 illustrates the comprehensive methodology adopted by the DTFWOA-ASI strategy.

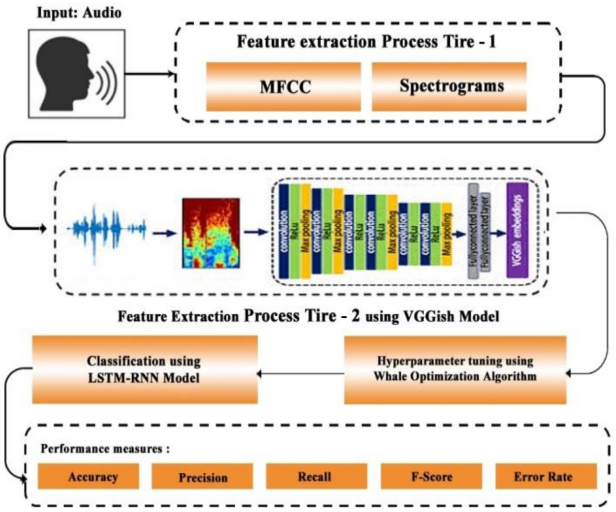
Input: Audio
Feature extraction Process Tire - 1
MFCC
Spectrograms
convolution
ReLu
Max pooling
Fullyconnected layer
VGGish embeddings
Feature Extraction Process Tire - 2 using VGGish Model
Classification using LSTM-RNN Model
Hyperparameter tuning using Whale Optimization Algorithm
Performance measures :
Accuracy
Precision
Recall
F-Score
Error Rate

Fig. 1. Overall Architecture of the DTFWOA-ASI Approach

3.1. Median filter-based noise removal

The median filter (MF) is renowned for its ability to maintain an image's sharpness while effectively removing noise. It replaces each pixel with the median value from its adjacent pixels, employing a 3x3 window for this purpose [22]. Among the conventional filters, it stands out as the most efficient in eradicating speckle noises. The process for developing the AMF is delineated in Algorithm 1. A key aspect is spatial processing, which ensures edge details are preserved and non-impulsive noises are eliminated via the adaptive MF. The upgraded model, known as the AMF, protects the detailed structures and contours within an image by adjusting the window size uniquely for each pixel.

Algorithm 1: Pseudocode of AMF

```
(1) Start by taking matrix “A”, which contains N
    columns and M rows as input.
(2) Enhance the input matrix by adding a zero on its
    peripheries, resulting in a new matrix with N + 2
    columns and M + 2 rows.
(3) Utilize a 3 × 3 size mask.
(4) Initially, position the mask over the upper left corner,
    specifically the first column and row of the “A”
    matrix.
(5) For every element covered by the mask, organize
    them in order from lowest to highest.
(6) The median value, or the middle value, from this
    ordered list, is then used to replace the element at
    position A (1, 1).
(7) Proceed by moving the mask to the next element.
(8) Repeat steps 4 through 7 until every element within
    matrix “A” has been updated with its corresponding
    median value.
```

3.2. Spectrograms and MFCC

In this study, we've employed Spectrograms alongside Mel Frequency Cepstral Coefficients (MFCC) as the core features for speech analysis. Renowned for its effectiveness in a myriad of speech-related applications, MFCC's base lies in the cepstral analysis of speech signals [23]. This process splits the speech signal into segments that lasts for 25 ms, with a shift of 10 ms between each. Afterward, these segments undergo application to a Hamming window, followed by a Fast Fourier Transform (FFT). The output from the FFT is then directed through a Mel-scale triangular filter bank. The outputs from this filter bank then undergo further processing via MFCC, leading to the derivation of Discrete Cosine Transform features. A critical step includes the normalization of MFCC by making adjustments for mean and variance, to standardize the spectrogram features to achieve zero mean and unit variance. This standardization is accomplished through the analysis of the speech signal in the frequency domain, utilizing a Hamming window that functions with a 10 m step size and a segment length of 25 ms. The spectrogram is then normalized for mean and variance, ensuring the features possess zero mean and unit variance.

3.3. VGGish-based Feature Extraction

The approach introduced employs the MFCC spectrogram for the Convolutional Neural Network's (CNN) input; this spectrogram, a bi-dimensional signal, encapsulates the speaker's unique identity markers. Concurrently, CNN ensures temporal and spatial translation is invariant, allowing for the capture of voiceprint characteristics within the spectrogram domain without halting the temporal sequence [24]. Consequently, the research propounds the utilization of both MFCC and the spectrogram as inputs into the VGGish system. Extracting audio features from sounds, particularly for purposes such as sound classification, event detection in sounds, and identifying speakers, involves a technique that leverages a method based on VGGish. This approach employs a pre-developed deep learning model titled VGGish, rooted in the VGG architecture, which is prevalently applied in categorizing images. Tailored for the extraction of features from audio files, VGGish is a variant of the convolutional neural network (CNN) setup [25]. It includes multiple convolutional layers succeeded by max-pooling stages, mirroring the structure used for image classification. Through this structure, VGGish transforms audio inputs into concise, fixed-length embeddings. These embeddings encapsulate substantial audio characteristics like pitch, timbre, and spectral properties in a succinct form as shown in Fig 2. Originally trained on an extensive compilation of audio snippets, the VGGish model masters universal sound representations. These learned weights are then utilized in extracting attributes from unfamiliar audio files. Before the processing by VGGish, audio signals are converted into mel-spectrogram formats, which is a technique for depicting time-frequency information of sound, especially effective in highlighting spectral characteristics.

The essence of audio signals is distilled into a 128-dimensional feature vector by the VGGish model. Learned across convolutional and pooling layers specifically trained on a vast array of audio segments, this feature vector encases high-level attributes of the sound input. Each of the 128 dimensions within the vector captures a unique aspect of the sound, encompassing elements such as spectral attributes, timbre, and pitch, presenting a compact yet comprehensive snapshot of the audio. Although pinpointing the precise significance of each dimension in the feature vector might be challenging, this representation becomes a valuable asset for several sound analysis purposes, including but not limited to classification, detection, and speaker recognition. With a standardized length (128 dimensions), the feature vector ensures swift processing and uniform representation for auditory inputs of varying sizes. This consistency renders the VGGish-based extraction method highly effective for numerous audio analytical activities, offering a condensed yet insightful portrayal of sound content for a multitude of analytical tasks.

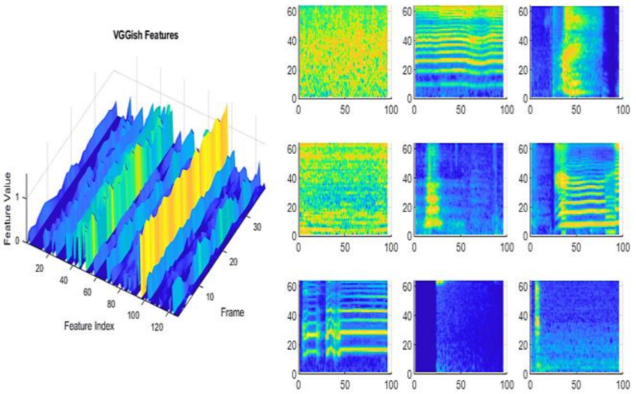

Fig. 2. Extraction of Sample VGGish Features

3.4. WOA-based Hyperparameter Tuning

Developed by Seyedali Mirjalili and colleagues in the year 2016, the Whale Optimization Algorithm draws its inspiration from the intricate social behaviors and hunting tactics employed by humpback whales [26]. This metaheuristic optimization algorithm is crafted to identify the most favorable solutions within continuous search spaces. Its application spans a wide array of optimization challenges.

- ***Initialization Phase (Initialize Whales):***

  Begin by setting up a group of candidate solutions, which act as potential answers to the optimization challenge. These solutions are similar to whales within the algorithm context. Every whale gets represented through a parameter vector within the search domain.

- ***Exploration Phase:***

  In the initial exploration phase, whales venture across the search area to identify areas full of potential. Whales adjust their locations by employing a strategy based on randomness.

- ***Exploitation Phase:***

  During the exploitation phase, whales become attracted to promising solutions identified in the exploration stage. They shift their positions in the direction of the globally best solution that has been discovered thus far.

- ***Position Update:***

  Designate the location of each whale in the sequence, known as X_i(t), at any given iteration t, with i representing the specific whale from a total count of N whales, and t marking the iteration as X_i(t), where i=1,2, ..., N (number of whales). For every whale, signified as i which ranges from 1 to N, symbolizing the total whales counted. The term t reflects the ongoing iteration.

The Eq. (1) is for updating positions within the Whale Optimization Algorithm are characterized as below:

$$X_i(\mathrm{t}+1) = \begin{cases} X_i(t) - A\ .\ D & if\ r_1 < 0.5 \\ X_{rand}(t) - A\ .\ dist(X_i(t), X_{rand}(t)) & otherwise \end{cases} \quad (1)$$

Where:

A is the amplitude parameter controlling the search step size.

D is the distance to the current best solution.

r_1 is a random number in the range [0, 1].

$X_{rand}(t)$ is a randomly selected whale position.

$dist(X_i(t), X_{rand}(t))$ is the Euclidean distance between $X_i(t)$ and $X_{rand}(t)$.

- ***Boundary Constraints:***
  Verify that the revised locations of the whales stay inside the acceptable searching area.
- ***Termination Phase:***
  The process of optimization comes to an end once it fulfills a specific criterion for termination. This could involve the process reaching its maximum allowed iterations or when the quality of the solution is deemed satisfactory.

Within the Whale Optimization Algorithm (WOA), it's the role of the fitness function to assess the caliber of each possible solution, or "whale," within the searching arena. This fitness function varies based on the problem it's addressing and relies on the optimization task's goal. In the realm of speaker identification tasks, the fitness function is typically tasked with gauging the effectiveness of a specific hyperparameter combo for the classifier in accurately pinpointing speakers.

Define *f(X)* as the evaluation metric, where in *X* symbolizes the array of hyperparameters (such as the structure of the classifier, the rates of learning, regularization parameters, and so forth). The evaluation metric, *f(X),* can be characterized by either accuracy or another appropriate measure that assesses how well the classifier, configured with a specific hyperparameters array, performs on a test dataset illustrated in Eq. (2).

$$f\ (X) = Accuracy \quad (2)$$

3.5. Classification using LSTM- RNN

Utilizing the LSTM-RNN architecture, it serves as a classifier in the automated ASR system. This model comprises an input layer, followed by two hidden layers, and culminates with an output layer. Typically, one could describe a feedforward neural network (FFNN) in the following manner [27].

$$Y = F(\ X, \theta\ ) \quad (3)$$

Where $X = \{x_1, x_2, \ldots, x_n\}$ refers to an input set, $Y = \{y_1, y_2, \ldots, y_m\}$ denote the output collection, while *F* signifies an FFNN unit, and θ is indicative of the unit's parameter compilation. Within a classification framework, *Y* symbolizes a classification group as illustrated in Eq. (3). A Convolutional Neural Network (CNN) forms a subset of FFNN and finds its use in semantic segmentation, classifying images, and identifying targets. The distinctive feature of the CNN approach lies in its engagement of convolution and pooling stages, distinguishing it from other NN frameworks. These convolution stages are tasked with the extraction of local attributes from the provided dataset.

$$Y_F = Conv\ (X\ ,\ \theta_{CONV}\ ) \quad (4)$$

A *Conv* represents a singular convolutional layer, and *YF* denotes the subset of features that the convolutional layer extracts from *X*. In this layer, θ_{CONV} comprises the set of parameters as shown in Eq. (4). The purpose of the pooling layer is to condense local features, thereby emphasizing the importance of the feature.

$$Y_{CF} = Pool\ (Y_{Fr}\ ,\ \theta_{Pool}\ ) \quad (5)$$

Poo1 signifies a pooling layer. YCF is the terminology used for a group of condensed features, and we demonstrate the blend between CNN and the pooling layer deriving from *YF*. θ_{Pool} is the collection of parameters in the pooling layer as illustrated in Eq. (5).

$$Y = Softmax\left(FC(Pool(Conv(X\ ,\ \theta_{CONV}\ )\ ,\ \theta_{Pool}\ ),\ \theta_{FC}\ )\right) \quad (6)$$

Regarding a module for classification, it incorporates a CNN consisting of FC and Softmax layers, and coupled with the anterior part of an RNN, it establishes a CRNN framework. The equation *Y = F (X, θ)* serves the function of classifying the features as shown in Eq. (6).

FC is an abbreviation for a fully connected layer, while SoftMax represents a Softmax layer. An RNN stands as a distinct type of FFNN, primarily utilized in datasets possessing a sequential structure, for instances like speech recognition and machine translation, to name a few. Among the RNN techniques, LSTM-RNN is particularly prevalent due to its capacity to tackle the vanishing gradient dilemma through the use of memory cells that retain long-duration data. As a tactic for classification, the LSTM-RNN comprises both Softmax and FC layers. The *Y = F (X, θ)* of LSTM – RNN is as shown in Eq. (7).

$$Y = Softmax\left(FC(LSTM(X\ ,\ \theta_{LSTM}\ ),\ \theta_{FC}\ )\right) \quad (7)$$

The methodology we embrace employs a singular-output LSTM-RNN for the process of decision-making as delineated in Fig 3. Our application involves fixed-length inputs despite LSTMs not being bound by such constraints.

The utterances transform feature sets, serving as input for the LSTM-RNN. A frame of 25 ms is utilized. The amount of input coefficients matches the input layer size of the LSTM-RNN. Contrary to functioning as a frame classifier, the network operates as a sequence classifier.

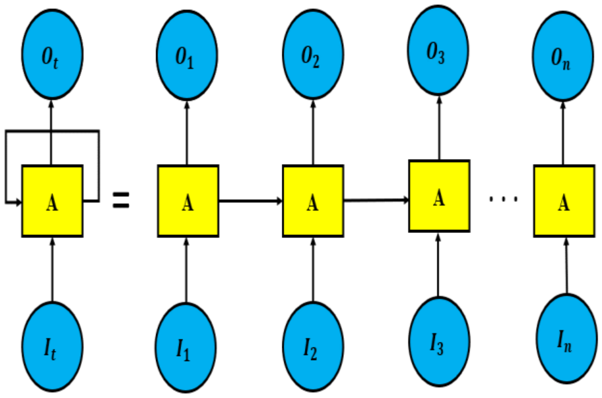

Fig. 3. Architecture of LSTM – RNN

The process of making decisions taps into the use of a loss function alongside the ultimate hidden state found within the LSTM architecture. This model features a dual-layer structure, each layer being equipped with 300 nodes. Following this, a SoftMax layer gets integrated, mirroring the number of classes available. Within the hidden layers of the LSTM, memory blocks serve the purpose of retaining the current condition of the network by acting as a storage device. Subsequently, a SoftMax layer, identical in class count, is appended. The probability that a specific frame is attributed to a certain speaker is determined by the SoftMax layer output, though this likelihood isn't solely based on the current frame but also considers every frame that preceded it in the sequence. The system is capable of making out each output by taking into account inputs from both past and present. Upon obtaining access to the full file, it proceeds to compute outcomes.

4. Experimental Results

4.1. Implementation Setup and Evaluation Metrics

Within this segment, the DTFWOA-ASI model underwent a series of experimental validations to identify speakers through audio file analysis, considering a variety of aspects. The testing phase leveraged Python 3.6.5 on a system configured with an i5-8600K CPU, a 250GB SSD, a GeForce 1050Ti 4GB graphics card, 16GB of memory, and a hard disk of 1TB capacity. For validation, a benchmark Kaggle dataset comprising audio files was utilized [28]. The total samples for the test files have been documented in Table 1. Evaluating the DTFWOA-ASI model's efficacy included essential metrics like Accuracy, Precision, Recall, F-score, and Error Rate as shown in Eq. (8-11). When the model accurately identifies the positive category, this is known as a True Positive (TP). Conversely, a True Negative (TN) is observed when the negative class is correctly identified by the model. Instances where the model inaccurately predicts the positive group are labeled as False Positives (FPs). Similarly, when the model incorrectly identifies the negative group, these are termed False Negatives (FNs). The description of these measures can be established as follows:

$$Accuracy = \frac{\mathrm{TP}+\mathrm{TN}}{TP+FP+FN+TN} \times 100 \qquad (8)$$

$$Precision = \frac{\mathrm{TP}}{TP+FP} \qquad (9)$$

$$Recall = \frac{\mathrm{TP}}{TP+FN} \qquad (10)$$

$$F - Score = \frac{2 * Precision * Recall}{Precision + Recall} \qquad (11)$$

Error Rate: An error rate is defined as the fraction of wrong answers observed in a method, network, or evaluation. It is often represented as either a percentage or a ratio, which contrasts the count of mistakes with the overall amount of observations or trials conducted.

Table 1. Dataset Details

Class	Number of Samples
Speaker 1	1500
Speaker 2	1500
Speaker 3	1500
Speaker 4	1500
Speaker 5	1500
Total	**7500**

Fig. 4 reveals the performance outcomes of the DTFWOA-ASI strategy with a 70:30 TR/TS dataset. From the data, it's evident that the DTFWOA-ASI method marked the highest training and validation accuracy figures. Remarkably, it should be noted that the accuracy obtained through testing seems to exceed what is achieved during the training phase. Additionally, the figures for training and validation loss produced by the DTFWOA-ASI approach, with the same 70:30 TR/TS division, are depicted in Fig. 4. The results suggest that the DTFWOA-ASI algorithm has been successful in reaching the lowest scores for both training loss and validation loss. Interestingly, the validation loss was demonstrated to be smaller than the training loss.

Table 2. Analyzing the Outcomes of the DTFWOA-ASI Method Across Various Metrics a 70:30 Split of the TR/TS Data

Labels	Accuracy (%)	Precision (%)	Recall (%)	Error Rate (%)	F-Score (%)
Training Phase (70%)					
Speaker – 1	94.64	85.19	86.79	05.36	85.98
Speaker – 2	93.57	86.27	80.00	06.43	83.02
Speaker – 3	96.43	93.48	86.00	03.57	89.58
Speaker – 4	96.79	88.73	98.44	03.21	93.33
Speaker – 5	95.71	89.66	89.66	04.29	89.66
Average	95.43	88.67	88.18	04.57	88.31
Testing Phase (30%)					
Speaker – 1	98.95	95.67	97.30	02.25	97.73
Speaker – 2	98.85	96.00	97.00	02.15	97.20
Speaker – 3	98.75	95.10	97.00	02.58	97.53
Speaker – 4	98.85	96.89	98.10	02.19	97.12
Speaker – 5	98.75	96.44	97.27	02.00	97.40
Average	98.83	96.02	97.33	01.05	97.39

Fig. 5 showcases the confusion matrices, Precision-Recall, and ROC curves generated by the DTFWOA-ASI model across various sizes of Training (TR) and Testing (TS) data. It's illustrated that the DTFWOA-ASI model has outperformed in recognizing speakers. Table 2 depicts the overall performance in speaker identification of the DTFWOA-ASI model, utilizing 70% training data and 30% testing data. The outcomes demonstrate the model's capability to accurately identify each category.

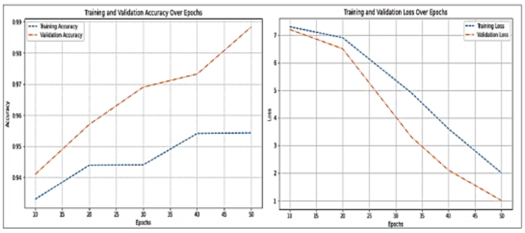

Fig 4. Accuracy & Loss graph based on training and testing set

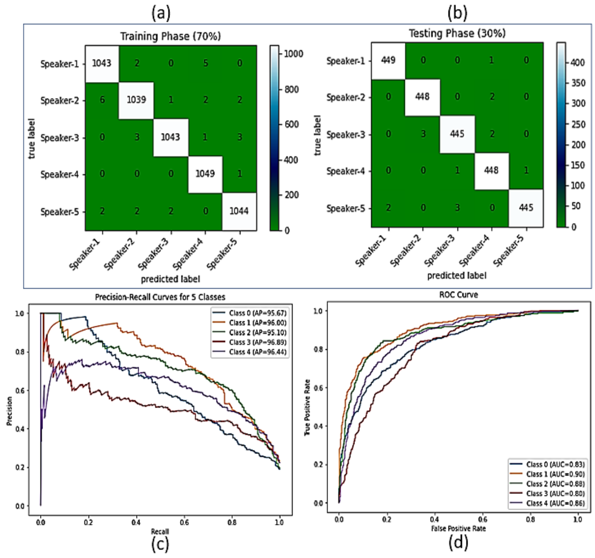

Fig. 5. (a) Confusion Matrix based on TR set (b) Confusion Matrix based on TS set (c) Precision-Recall Curve (d) ROC Curve

To showcase the enhanced performance of the DTFWOA-ASI configuration, a comparative analysis was conducted. The outcomes of this study are depicted in Table.3 [11,27]. Additionally, Fig. 6 offers an accuracy comparison between the DTFWOA-ASI strategy and other contemporary frameworks. The examination shows that the framework combining MFCC-SOFM-MLP-GD recorded the lowest success rate, hitting a 96.92% mark. On the other hand, slightly improved performances were observed with the MFCC-SOFM-MLP-GDM, MFCC-SOFM-MLP-BR, MFCC-FW, and the fusion methods, which respectively achieved accuracy scores of 97.05%, 97.62%, 97.32%, and 97.81%. However, the proposed DTFWOA-ASI approach that demonstrated the most impressive outcomes, achieving an elevated accuracy of 98.83%.

Table. 3. Evaluating the Accuracy of the DTFWOA-ASI Method Versus Contemporary Techniques

Models	Accuracy (%)	Error rate (%)
MFCC-SOFM-MLP-GD	96.92	03.08
MFCC-SOFM-MLP-GDM	97.05	02.95
MFCC-SOFM-MLP-BR	97.62	02.38
MFCC-FW	97.32	02.68
FUSION	97.81	02.19
Proposed DTFWOA-ASI	**98.83**	**01.05**

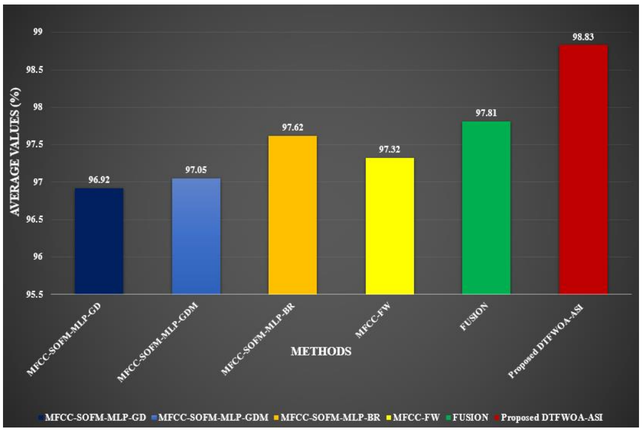

Fig. 6. Evaluation of an Accuracy of the DTFWOA-ASI Framework with existing methodologies

5. Conclusion

In the study documented here, a pioneering DTFWOA-ASI framework was constructed to authenticate speaker identification applications effectively. This cutting-edge DTFWOA-ASI mechanism initially applies the AMF approach for eradicating noise interferences in audio signals. Subsequently, it incorporates MFCC and spectrograms as input for the VGGish architecture. Following this, the Whale Optimization Algorithm (WOA) is deployed for refining the hyperparameters connected to the LSTM-RNN design. In the final stage, the LSTM-RNN architecture is harnessed as a classifier to facilitate automated speech recognition (ASR). The evaluation of the DTFWOA-ASI framework's efficacy was executed through a comprehensive series of tests. A comparative analysis underscored the superior performance of the DTFWOA-ASI framework over other contemporary methods, establishing its potential for robust ASR in real-time Speaker Identification scenarios. Looking ahead, a combined approach employing fusion-based deep learning models might be explored to enhance the DTFWOA-ASI framework's performance further.

References

[1] Machado, T.J.; Vieira Filho, J.; de Oliveira, M.A. Forensic speaker verification using ordinary least squares. Sensors 2019, 19, 4385.

[2] Wang, Z.; Xia, W.; Hansen, J.H. Cross-domain adaptation with discrepancy minimization for text-independent forensic speaker verification. arXiv 2020, arXiv:2009.02444.

[3] Stefanus, I.; Sarwono, R.J.; Mandasari, M.I. GMM-based automatic speaker verification system development for forensics in Bahasa Indonesia. In Proceedings of the 2017 5th International Conference on Instrumentation, Control, and Automation (ICA),Yogyakarta, Indonesia, 9–11 August 2017; pp. 56–61.

[4] Algabri, M.; Mathkour, H.; Bencherif, M.A.; Alsulaiman, M.; Mekhtiche, M.A. Automatic speaker recognition for mobile forensic applications. Mob. Inf. Syst. 2017, 2017, 6986391.

[5] Gaurav, B.S.; Agarwal, R. An efficient speaker identification framework based on Mask R-CNN classifier parameter optimized using hosted cuckoo optimization (HCO). J. Ambient Intell. Human. Comput. 2022, 13, 1–13.

[6] Susanto, S.; Wang, Z.; Wang, Y.; Nanda, D.S. Forensic Linguistic Inquiry into the Validity of F0 as Discriminatory Potential in the System of Forensic Speaker Verification. J. Forensic Sci. Crim. Investig. 2017, 5, 555664.

[7] Nagrani, A.; Chung, J.S.; Xie, W.; Zisserman, A. Voxceleb: Large-scale speaker verification in the wild. Comput. Speech Lang. 2020, 60, 101027.

[8] Athulya, M.S.; Sathidevi, P.S. Speaker verification from codec distorted speech for forensic investigation through serial combination of classifiers. Digit. Investig. 2018, 25, 70–77.

[9] Hautamäki, R.G.; Sahidullah, M.; Hautamäki, V.; Kinnunen, T. Acoustical and perceptual study of voice disguise by age modification in speaker verification. Speech Commun. 2017, 95, 1–15.

[10] Das, R.K.; Prasanna, S.M. Speaker verification from short utterance perspective: A review. IETE Tech. Rev. 2018, 35, 599–617.

[11] Susanto, S.; Nanda, D.S. December. Analyzing Forensic Speaker Verification by Utilizing Artificial Neural Network. In International Congress of Indonesian Linguistics Society (KIMLI 2021); Atlantis Press: Amsterdam, The Netherlands, 2021; pp. 128–132.

[12] Al-Ali, A.K.H.; Dean, D.; Senadji, B.; Chandran, V.; Naik, G.R. Enhanced forensic speaker verification using a combination of DWT and MFCC feature warping in the presence of noise and reverberation conditions. IEEE Access 2017, 5, 15400–15413.

[13] Huang, S.; Dang, H.; Jiang, R.; Hao, Y.; Xue, C.; Gu, W. Multilayer Hybrid Fuzzy Classification Based on SVM and Improved PSO for Speech Emotion Recognition. Electronics 2021, 10, 2891.

[14] Swain, M.; Maji, B.; Kabisatpathy, P.; Routray, A. A DCRNN-based ensemble classifier for speech emotion recognition in Odia language. Complex Intell. Syst. 2022, 8, 4237–4249.

[15] Mardhotillah, R.; Dirgantoro, B.; Setianingsih, C. Speaker Recognition for Digital Forensic Audio Analysis using Support Vector Machine. In Proceedings of the 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 10–11 December 2020; pp. 514–519.

[16] Saleem, S.; Subhan, F.; Naseer, N.; Bais, A.; Imtiaz, A. Forensic speaker recognition: A new method based on extracting accent and language information from short utterances. Forensic Sci. Int. Digit. Investig. 2020, 34, 300982.

[17] Khan, F.; Tarimer, I.; Alwageed, H.S.; Karada ˘g, B.C.; Fayaz, M.; Abdusalomov, A.B.; Cho, Y.-I. Effect of Feature Selection on the Accuracy of Music Popularity Classification Using Machine Learning Algorithms. Electronics 2022, 11, 3518.

[18] Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5329–5333.

[19] NIST. Speaker Recognition Evaluation 2016. Available online: https://www.nist.gov/itl/iad/mig/speaker-recognition evaluation-2016/ (accessed on 30 July 2020).

[20] Devi, K.J.; Singh, N.H.; Thongam, K. Automatic speaker recognition from speech signals using self-organizing feature map and hybrid neural network. Microprocess. Microsyst. 2020, 79, 103264.

[21] Teixeira, F.; Abad, A.; Raj, B.; Trancoso, I. Towards End-to-End Private Automatic Speaker Recognition. arXiv 2022, arXiv:2206.11750.

[22] Gao, H.; Hu, M.; Gao, T.; Cheng, R. Robust detection of median filtering based on combined features of the difference image.Signal Process. Image Commun. 2019, 72, 126–133.

[23] Ma, Z.; Fokoué, E. Accent Recognition for Noisy Audio Signals. Serdica J. Comput. 2014, 8, 169–182.

[24] Wang, C.; Chen, D.; Hao, L.; Liu, X.; Zeng, Y.; Chen, J.; Zhang, G. Pulmonary image classification based on inception-v3 transfer learning model. IEEE Access 2019, 7, 146533–146541.

[25] Tsalera, Eleni, Andreas Papadakis, and Maria Samarakou. "Comparison of pre-trained CNNs for audio classification using transfer learning." Journal of Sensor and Actuator Networks 10.4 (2021): 72.

[26] Gowrishankar, Bettadamadahally Shivakumaraswamy, and Nagappa U. Bhajantri. "Raga classification using enhanced spatial bound whale optimization algorithm." Indonesian Journal of Electrical Engineering and Computer Science 30.2 (2023): 825.

[27] Zhang, Y.; Xiong, R.; He, H.; Pecht, M.G. Long short-term memory recurrent neural network for remaining useful life prediction of lithium-ion batteries. IEEE Trans. Veh. Technol. 2018, 67, 5695–5705.

[28] https://www.kaggle.com/code/auishikpyne/speaker-identification/ input