Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.059382

ARTICLE

Enhancing User Experience in AI-Powered Human-Computer Communication with Vocal Emotions Identification Using a Novel Deep Learning Method

Ahmed Alhussen[1], Arshiya Sajid Ansari[2,*] and Mohammad Sajid Mohammadi[3]

[1]Department of Computer Engineering, College of Computer and Information Sciences, Majmaah University, Al-Majmaah, 11952, Saudi Arabia

[2]Department of Information Technology, College of Computer and Information Sciences, Majmaah University, Al-Majmaah, 11952, Saudi Arabia

[3]Department of Computer Science, College of Engineering and Information Technology, Onaizah Colleges, Qassim, 51911, Saudi Arabia

*Corresponding Author: Arshiya Sajid Ansari. Email: ar.ansari@mu.edu.sa

Received: 06 October 2024 Accepted: 06 December 2024 Published: 17 February 2025

ABSTRACT

Voice, motion, and mimicry are naturalistic control modalities that have replaced text or display-driven control in human-computer communication (HCC). Specifically, the vocals contain a lot of knowledge, revealing details about the speaker's goals and desires, as well as their internal condition. Certain vocal characteristics reveal the speaker's mood, intention, and motivation, while word study assists the speaker's demand to be understood. Voice emotion recognition has become an essential component of modern HCC networks. Integrating findings from the various disciplines involved in identifying vocal emotions is also challenging. Many sound analysis techniques were developed in the past. Learning about the development of artificial intelligence (AI), and especially Deep Learning (DL) technology, research incorporating real data is becoming increasingly common these days. Thus, this research presents a novel selfish herd optimization-tuned long/short-term memory (SHO-LSTM) strategy to identify vocal emotions in human communication. The RAVDESS public dataset is used to train the suggested SHO-LSTM technique. Mel-frequency cepstral coefficient (MFCC) and wiener filter (WF) techniques are used, respectively, to remove noise and extract features from the data. LSTM and SHO are applied to the extracted data to optimize the LSTM network's parameters for effective emotion recognition. Python Software was used to execute our proposed framework. In the finding assessment phase, Numerous metrics are used to evaluate the proposed model's detection capability, Such as F1-score (95%), precision (95%), recall (96%), and accuracy (97%). The suggested approach is tested on a Python platform, and the SHO-LSTM's outcomes are contrasted with those of other previously conducted research. Based on comparative assessments, our suggested approach outperforms the current approaches in vocal emotion recognition.

KEYWORDS

Human-computer communication (HCC); vocal emotions; live vocal; artificial intelligence (AI); deep learning (DL); selfish herd optimization-tuned long/short K term memory (SHO-LSTM)

1 Introduction

Speech emotion recognition (SER) has garnered increasing attention in recent years as it utilizes speech cues to assess emotional states. SER is a challenging task, though it requires the identification of practical emotional components. SER comes in useful while researching computer-human identification. This means that to define the system's operations effectively, the system must understand the user's sentiments [1]. In interpersonal interactions, emotional intelligence is essential. Real-time implementation is prohibitively costly due to the technological difficulty of emotion identification using facial recognition. The implementation costs are also expensive as getting facial images needs high-quality cameras [2]. As science and technology advance, several innovators in the field are attempting to integrate text, vision, audio, and more multimodal information sources to progress the Human-Computer Interaction (HCI) technological sector. In both academics and business, multimodal interaction is currently quite popular [3]. Advances in AI-driven human-computer communication have concentrated on the recognition of vocal emotions through deep learning, which allows for the identification of subtle emotional signals such as tone and pitch. This will make interactions involving virtual assistants, customer care chatbots, and other related applications more empathetic and context-aware [4]. Digital people are embodied conversational agents that have been enhanced with AI. An artificially intelligent computer-based discussion system is known as an embodied conversational agent. To react to user input, digital beings employ audio-vocal communication methods and autonomous face animation [5]. When SER is used with a human-computer dialogue system, it can make the system intelligent and compassionate by recognizing the emotions expressed in speech, utilizing human voice as input, and enabling rich emotional communication between people [6]. Since the dynamic process of emotion recognition focuses on the emotional state of the individual, there are differences in the feelings that correlate to each person's activities. People use a variety of methods to express their emotions. It's critical to accurately perceive these emotions to facilitate effective conversation [7]. Numerous virtual agents and communicative robots have been developed in various sectors for a wide range of applications as robotics and AI technology have advanced. Because computer-graphics-based agents' programs can be loaded on popular devices like Personal Computers (PCs), tablets, and smartphones, they have the potential to be extensively used by many people. These agents also benefit from the ability to express themselves without being physically constrained, even when it comes to excessive emotional reactions [8,9]. Machine learning applied to human-computer communication processes involves the improvement of accuracy and efficiency. Such systems use large speech datasets and sophisticated neural networks to accurately detect and interpret diverse emotions, which they use to enhance interactions with applications such as smart assistants and customer support applications that respond according to the emotional context [10]. In emotion recognition techniques, integration of real-time adaptive models is highlighted to improve emotional responses in a more nuanced and contextually relevant human-computer interaction [11]. In addition, recent developments in deep learning techniques have considerably made vocal emotion recognition more accurate and efficient, thus promising more personalized and empathetic responses in user interactions with AI-based applications such as virtual assistants and chatbots [12].

The objective is to generate and evaluate a novel AI system selfish herd optimization-tuned long/short-term memory (SHO-LSTM) strategy to identify vocal emotions in human communication.

Contribution of the Study

- We collected the RAVDESS dataset and preprocessed the data using the wiener filter method.
- The feature selection process uses MFCC, making selections from the audio data in the Vocal Emotions Identification.

- Then, we introduce an innovative method called the SHO-LSTM strategy to identify vocal emotions in human communication.
- The paper described the SHO-LSTM approach to improve the training efficiency of human emotions.

The structure of the article is described as follows: Section 2 has a literature review. Section 3 provides a detailed methodology. Section 4 gives an analysis of the findings and Section 5 delivers a conclusion.

2 Related Works

A Study [13] described that emotions play a vital role in our mental functioning. They are essential to determining the conduct and mental state of an individual. The process of identifying Speech emotion recognition (SER) is nothing but the process of identifying a speaker's emotional state from their speech signal. For classification, two suggested models were used: a special 2D Convolutional neural network (2D CNN) architecture in addition to a 1D CNN with LSTM and focus integrated. The results showed that the indicated 1D CNN outperformed the 2D CNN using LSTM and concentration. Article [14] proposed a framework that enables seamless collaboration between people and sophisticated computer systems, such as robots. To be possible, the computer system must be able to transmit information in some way to others. A feature that allows a system to comprehend human emotions and communicate those feelings to human counterparts must be included in the system. Investigation [15] suggested an approach to sentiment identification using Electroencephalogram (EEG) signals and transfer support vector machines (TSVM). The need for human-computer interaction systems (HCIS) to be intelligent is currently becoming more and more apparent. They developed the heuristic multimodal real-time emotional identified technique (HMR-TER) to offer immediate and appropriate feedback via the Internet to learners based on their facial movements and vocal sounds [16]. The ability to recognize human emotion was crucial in interpersonal connections. Emotions are reflected in speech, hand and body movements, and facial expressions. Paper [17] determined an intelligent speech emotion detection model based on the feature representation of a Convolution CNN, was developed to increase the accuracy of the system. Speech recognition technology was used to identify different emotion kinds from provided attribute segments. A popular area of research in the speech sector was the creation of high-accuracy voice emotion detection systems, as the need for such systems grows in the corporate, educational, and other domains. They described the emotion recognition module integrating speech to achieve multimodal identification of emotions in [18]. CNN and LSTM algorithms were used to mine the textual and auditory data, which can help to obtain further the subjective content that was there. In the meantime, the sensation reporting mechanism uses the identified emotions to determine what psychological feedback is acceptable. Study [19] aimed to analyze and categorize the AI techniques, algorithms, and sensor technologies used in current human-computer intelligent interaction (HCII) research to classify the available data, suggest possible future research avenues and investigate trends in HCII research. AI methodologies, algorithms, and sensor technologies provide the foundation for intelligent solutions that enable computers to communicate with humans and function naturally. The suggested deep residue shrinking networks with the bi-directional gated recurrent unit (DRSN-BiGRU) technique consist of a convolution system, a residual shrinking system, a bidirectional recurrent system, and a fully linked network. The process of SER is very important in HCI. In [20], the aspects and modeling

of SER are examined. The mel-spectrogram was explained in full, along with its theory and extraction procedure, and it was used as a characteristic of speech. Table 1 compares the prior studies.

Table 1: Literature survey

Reference	Methods	Finding	Advantages	Disadvantages	Limitation
[13]	1D CNN with LSTM and focus *vs.* 2D CNN.	1D CNN outperformed 2D CNN using LSTM and focus.	Higher accuracy in emotion recognition.	Complexity in model design and training.	May require significant computational resources.
[14]	Framework for human-computer collaboration with emotion communication.	Computer systems must comprehend and communicate human emotions.	Enhances human-computer interaction.	Implementation complexity.	Integration with existing systems can be challenging.
[15]	Sentiment identification using EEG signals and TSVM.	Developed HMR-TER for real-time emotion recognition.	Provides timely and pertinent feedback based on facial and vocal data.	Requires multimodal data processing.	Real-time implementation can be resource intensive.
[16]	The Heuristic multimodal real-time emotional recognition (HMR-TER).	Recognized human emotions based on facial movements and vocal sounds.	Real-time emotion recognition.	Requires accurate facial and vocal emotion data.	Dependent on the quality of input data.
[17]	Intelligent speech emotion detection using Convolution CNN.	Developed model based on CNN feature representation for high accuracy.	Increased accuracy in emotion detection.	Limited to speech data.	May not generalize well to other forms of emotion data.
[18]	Multimodal emotion recognition using CNN and LSTM.	Integrated speech and text for emotion recognition.	Combines textual and auditory data for better accuracy.	Complexity in model integration.	May require extensive training data.
[19]	Analysis of AI techniques, algorithms, and sensors in HCII.	Suggested DRSN-BiGRU for SER.	Comprehensive analysis of current techniques.	Complexity in implementing suggested models.	May not cover all potential future research avenues.
[20]	Aspects and modeling of SER using mel-spectrogram.	Detailed theory and extraction procedure of mel-spectrogram.	Provides an in-depth understanding of SER.	Limited to mel-spectrogram feature.	May not include other important features for SER.

2.1 Problem Statement

The HCI systems must be able to understand user emotions to provide a seamless and intuitive interaction experience. However, speech emotion recognition algorithms can miss the subtleties and complexity of human expression, which lowers user satisfaction and leads to poor communication outcomes. The task at hand involves creating a dependable and strong DL technique that can accurately and sensitively identify speech emotions, thereby improving user experience in AI-powered human-computer communication. This issue calls for the investigation of novel approaches that can

successfully handle the intricacies of emotional expression in speech signals, opening the door for artificial intelligence systems that are more sensitive and empathic.

The vocal emotions data was first collected for study and the dataset followed meticulous preprocessing using the wiener filter method. Next, we use MFCC for emotion feature selection. The suggested SHO-LSTM was utilized in the study to categorize emotions in human interactions. The suggested methodology's flow is displayed in Fig. 1.

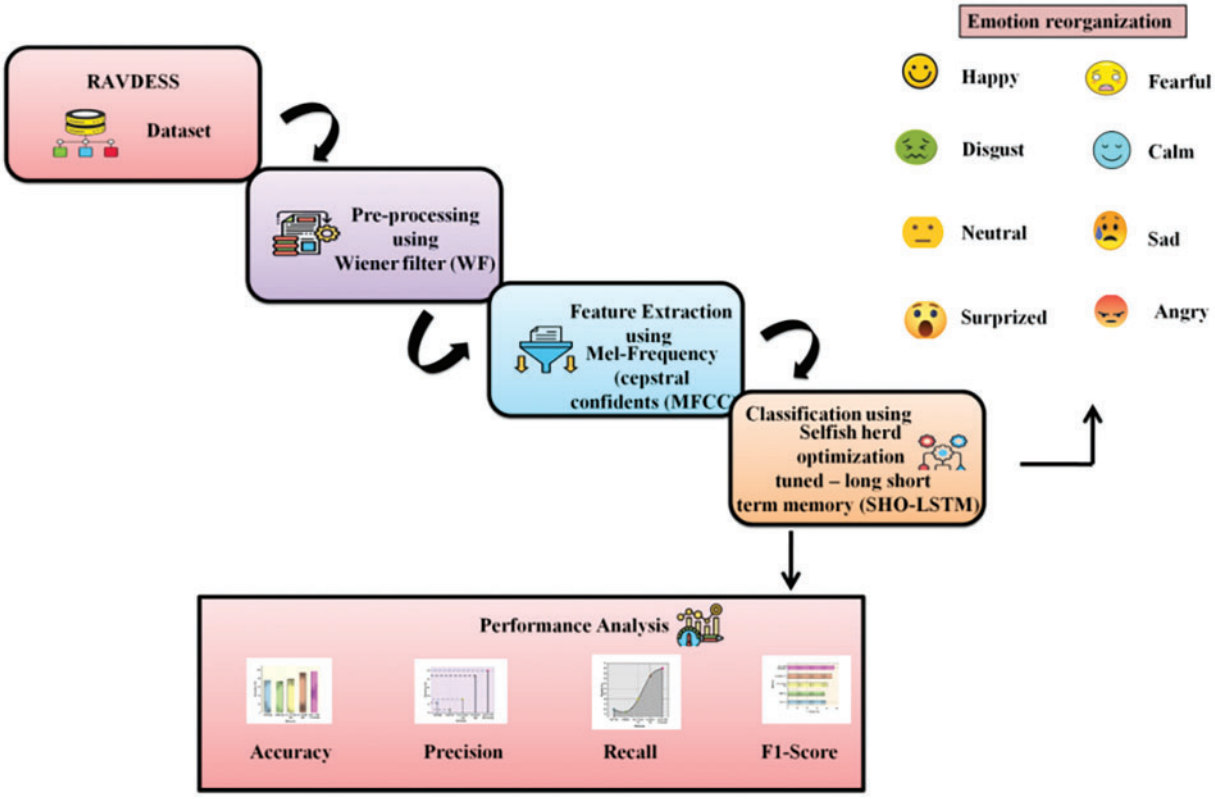

Figure 1: Structure of research framework

3 Methodology

3.1 Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset, which was collected via Kaggle (https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio, accessed on 12 September 2024), was used in this work. It consists of 1440 high-quality audio recordings of performers expressing eight emotions Happy, fearful, calm, disgusted, neutral, sad, surprised, and angry through song and speech as shown in Fig. 2, specifically for emotion detection research. Furthermore, the Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D) (https://www.kaggle.com/datasets/ejlok1/cremad, accessed on 12 September 2024), is a crowd-sourced emotional multimodal actor's dataset, with more than 7000 audio recordings of actors that express emotions under different speech and multimodal conditions. The dataset encompasses six basic emotions: anger, disgust, fear, happiness, sadness, and surprise, recorded from different speakers in both audio and visual modalities.

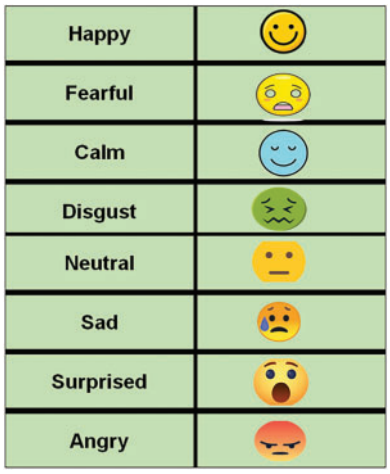

Figure 2: Emotions from the dataset

3.2 Preprocessing Wiener Filter (WF)

After collecting the data, a wiener filter is used to preprocess the collected data. The wiener filter provides popular audio processing techniques for removing unnecessary noise from multimodal vocal emotions. This technique produced a two-dimensional representation of the input image with values ranging from 0 to 255 for the pixels. The WF reduced mean square error and eliminated additive noise by using a linear estimate of the original image. This led to the computation of each pixel's mean and variance using Eqs. (1) and (2). The accuracy of speech emotion recognition therefore impacts the user experience in AI-powered human-computer interaction applications depending on this pre-processing stage.

$$\mu_{b,a} = \frac{1}{BA} \sum_{w,z \in M} O(W, Z) \tag{1}$$

$$\sigma^2{}_{b,a} = \frac{1}{BA} \sum_{w,z \in M} O(W, Z)^2 - \mu^2 \tag{2}$$

Here, O is an input image, W and Z are filtering pixels, $M = W = 3$ is surrounding pixels, and W and Z are placed in an $M \times X$ block.

$$O_{(b,a)} = \mu + \frac{\sigma^2 - m^2}{\sigma^2}(O\,(W - Z) - \mu) \tag{3}$$

In Eq. (3), where m^2 is the noise variance; adjustments are made. By using the WF approach we effectively reduced the noise in audio data, which is crucial for improving the clarity of vocal inputs for recognizing emotions is necessary to improve the user interaction with AI-powered human-computer systems of communication.

3.3 Employing Mel-Frequency Cepstral Coefficients (MFCC) to Extract Features

Following preprocessing, MFCC was used to select relevant audio features for vocal emotion identification to improve the user experience in an AI-powered human-computer interaction system. The Mel filter serves as the foundation for MFCC, one of the characteristic parameters that are frequently employed in voice recognition. The Mel filter's design considers the hearing system in

human ears. It takes into how people produce voices and the way human ears perceive them. The "Mel frequency," which refers to the frequency of the Mel filter is congruent with the auditory characteristics of the human ear and it is closely linked to the frequency of speech. The Mel frequency and frequency have a nonlinear connection. The precise correlation between the real frequency and the Mel frequency is as follows:

$$e_{mel} = 2595 \times kh\left(1 + \frac{e}{700}\right) \tag{4}$$

where e is the real speech frequency in hertz (H_z). The cepstral coefficient predicted from the Mel frequency is known as the MFCC. The process for obtaining MFCC is as follows:

- *Step 1: Initial Preprocessing (The Structuring)*

Following framing, the finite discrete signal ($W_j(m)$) is obtained, where j denotes the jth frame.

- *Step 2: Fast Fourier Transform (FFT)*

Every frame is given the FFT:

$$W(j,l) = EES\left[w_j(m)\right] = \sum\nolimits_{m=0}^{M-1} w_j(m)\, X_M^{lm}\; l = 0, 1, \cdots, M-1 \tag{5}$$

Here, M is how many sampling points are in every frame. The power spectrum is acquired $F(j,l) = [W(j,l)]^2$

- *Step 3: The Mel Filter Bank*

Use the Mel filter to calculate the energy.

$$t(j,n) = \sum\nolimits_{l=0}^{M-1} F(j,l)\, G_n(l) \tag{6}$$

- *Step 4: Discrete Cosine Transform (DCT)*

Use the DCT to calculate the MFCC.

$$MFCC(j,m) = \sqrt{\frac{2}{N}} \sum\nolimits_{n=0}^{N-1} \log[T(j,n)] \cos\left(\frac{\pi m(n-0.5)}{N}\right) m = 1, 2, \cdots K \tag{7}$$

where K is assumed to be 12 and m is the MFCC order. The number of Mel filters is N. The power spectrum is processed through a bank of Mel filters, which are triangle filters placed uniformly over the Mel scale and simulate Human hearing. This process accentuates crucial frequency bands while suppressing unimportant ones. It decreases the dimensionality of the feature space in comparison to raw audio, making it easier to manage and lowering computing complexity. This improvement of emotion recognition in a human-computer communication system mainly depends on these improvements' properties.

3.4 Vocal Emotions Identification Using Selfish Herd Optimization-Tuned Long\|Short-Term Memory (SHO-LSTM)

A novel Selfish Herd Optimization-tuned Long Short-Term Memory (SHO-LSTM), was developed to improve vocal emotion identification performance in AI-powered human-computer communication. Making use of the Selfish Herd Optimization algorithm for hyperparameter tuning, SHO-LSTM effectively captures and processes the subtle emotional nuance in speech, with the result

that the model's accuracy, as well as, interpretability is superior to traditional methods applied in an emotion recognition task.

3.4.1 Long Short-Term Memory (LSTM)

The long-term connections between sequential data can be comprehended by a sort of recurrent neural network called an LSTM, and therefore it might be more appropriate for the recognition of vocal emotions within AI-powered human-computer communication. The vocal emotions that enter cell memory are transformed into the final cell state and are preserved by the LSTM, as seen in Fig. 3. The input, forgets, update, and output barriers make up the major parts of the overall structure of the LSTM cell.

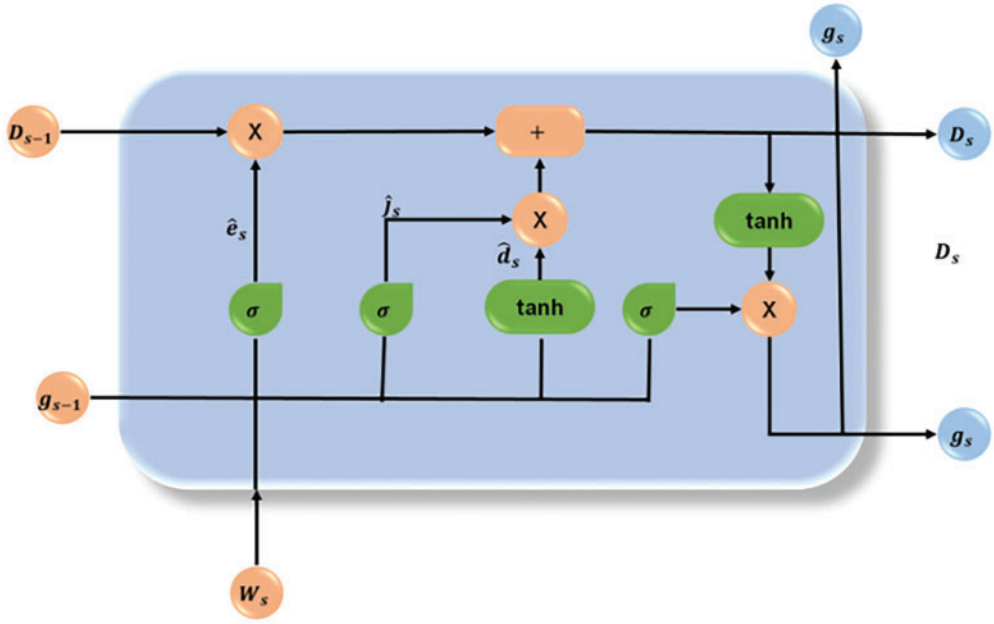

Figure 3: Architecture of LSTM

The information that has been received by earlier memory units is decided upon by the forget gate, accepted by the input gate, generated by the output gate into fresh long-term memory, and updated by the update gate into the cell. These four parts functions and interact in a certain way, accepting the audio input sequences of the LSTM memories at a certain time step and producing LSTM memories at the same time step. The input gate decides what vocal emotion needs to be sent to the cell and is expressed mathematically in the following equation:

$$j_i = \sigma(X_j * [g_{s-1,w_s}] + A_j) \tag{8}$$

The forget gate, which decides information is negated and expressed mathematically in the following equation:

$$e_j = \sigma(X_j * [g_{s-1,w_s}] + A_e) \tag{9}$$

The cell state is updated by the update gate and is described mathematically by the following equations:

$$D_s = tang(w_e * [g_{s-1,w_s}] + A_s) \tag{10}$$

$$D_s = e_s * B_{s-1,w_s} + j_s * B_s)$$

The output gate oversees updating the output, which is determined by the equation that follows. It oversees updating the preceding time step's hidden layer.

$$P_s = \sigma(X_o * [g_{s-1,w_s}] + A_e$$

$$G_s = P_s * tang(D_s) \tag{11}$$

LSTM effectively captures subtle voice emotions, promoting coherence and authenticity in communication dynamics by utilizing both short- and long-term memory.

3.4.2 Selfish Herd Optimization (SHO)

LSTM is a recurrent neural network type that understands the long-term dependencies across sequential data, and hence might be more suitable for the recognition of vocal emotions within AI-powered human-computer communication.

Analysis of Inspiration

To enhance the user experience in AI-powered human-computer communication, we integrate the Selfish Herd Optimization (SHO) approach to improve the performance of our Long Short-Term Memory (LSTM) model. The SHO algorithm was developed using the traits of selfish behavior exhibited by several animal groupings. It is considered that members of the group can converse with one another on the broad plain that is the search space. The SHO algorithm is based on the animal selfish herd hypothesis, which states that to increase its chances of surviving, every member of the herd would try to move in the direction of any potential predatory attack. However, the member won't consider its movement might impact the chances of other herd members surviving. SHO flow is shown in Fig. 4.

SHO's Mathematical Model

There are two groups within the SHO population: a herd of predators and a herd of prey, also known as the selfish herd. In the domain of SHO optimization, these two cohorts are considered search agents. Different evolutionary operators are applied according to distinct behavioral traits that are investigated in this predator-prey interaction. The eight phases that make up the SHO approach are listed below:

1. *Population Initialization:* The method begins by generating the positions of all animals as a single group, including predators (LSTM approach) and prey (representing the emotional identification), with the animals' locations limited to the lower bound (*lb*) and upper bound (*ub*).
2. *Group Separation:* The group is split into two groups prey (emotional detection technique) and predator (approaches that explore the solution) in the next phase. The size of the prey group may be calculated using the formula: $m_{prey} = floor(m \times rand\,(0.7, 0.9))$.

In this case, n is the overall population of the group, comprising both prey and predators, whereas m_{prey} is the number of individuals in the prey group.

3. *Calculating Fitness and Survival Rate:* Next, each member of the predator and prey group's survival rate is determined separately using (12), assessing the fitness of each model by determining how well it can recognize voice emotion.

$$TQ = \frac{e\left(z_j\right) - e_{best}}{e_{worst} - e_{best}} \tag{12}$$

Here, z_j indicates the fitness related to the position of the prey or predator, while e_{best} and e_{worst} stand for the best and worst fitness values discovered thus far.

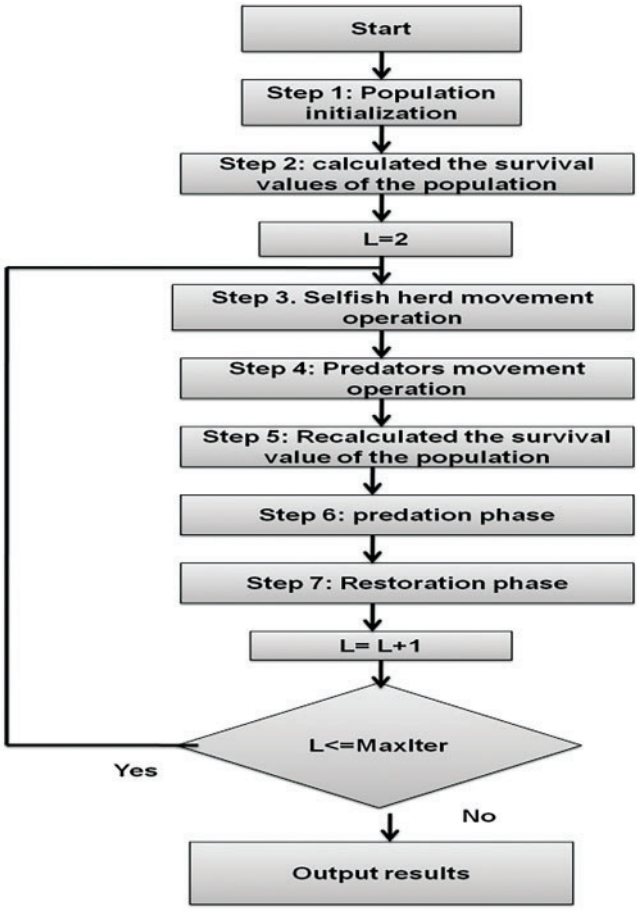

Figure 4: SHO flow

4. *Movement Update:* As it deals with the movement of every herd member, this stage is the most crucial in SHO. The prey model (high-performing LSTM) updates their position toward optimal solutions, while predators explore new areas to avoid overfitting. The following formula is applied to update the herd leader's (hl) location:

$$g_k = \begin{cases} g_k + 2 \times \varphi K.on \times (o_n - g_n) \; if \; TU_{gK} = 1 \\ g_k + 2 \times \alpha \times \varphi K.w_{best} \times (o_{best} - g_K) \; if \; TU_{g_K} < 1 \end{cases} \tag{13}$$

where α is a random value in the interval [0, 1] and φK is the self-serving repulsion that the current herd's leader is exhibiting toward the predators' center of Mass *on*. The parameter φK signifies the self-serving pull that the herd's leader, g_k, experiences in the direction of the global top position, w_{best}.

Moreover, it has two choices for updating the location of a herd member g_k: it may either follow the herd as indicated by Eq. (14) or abandon the group as shown by Eq. (15). This decision is based on a random variable.

$$g_j = g_j + e_j \tag{14}$$

where the definition of e_j is:

$$e_j = \begin{cases} 2 \times \left(\beta \times \psi_{g_j}, g_k \times \left(g_K - g_j\right) + \gamma \times \psi_{g_j, g_d} \left(g_d - g_j\right)\right) TU_{g_K} \leq TU_{g_\mu} \\ 2 \times \delta \times \psi_{g_j, g_n} \times \left(g_N - g_j\right) \text{otherwise} \end{cases} \tag{15}$$

Here, φ, β, and δ each represent a random number in the interval [0, 1], whereas ψ_{g_j}, g_k and ψ_{g_j, g_n} represent the selfish attraction that herd member hi experiments with toward g_k and g_d.

$$g_j = g_j + 2(\beta \times \psi_{g_k, w_{best}} \times \left(w_{best} - g_j\right) + \gamma \times (1 - TU_{g_j}) \times \hat{q}) \tag{16}$$

where $\hat{q}$ is the unit vector inside the n-dimensional search space that points in a random direction.

5. *Phase of Predation (Overfitting Risk):* Penalizing overfitting models ensures robust emotion recognition without memorizing noise. The predator's movement is then computed based on the pursuit probability, which is described as:

$$o_j = \frac{\omega_{oj}, i_i}{\sum_{n=1}^{Mg} \omega_{oj}, i_i} \tag{17}$$

The prey attraction between o_j and g_j is shown by ω_{oj}, i_i. The predator x p's location is updated using

$$w_o = w_o + 2 \times \rho \times (g_q - w_o) \tag{18}$$

In this case, ρ represents a random number within the range [0, 1].

6. Additionally, using Eq. (13), the survival rate of each member of the predator and prey group is computed separately.

7. *Mating Process (Phase of Restoration):* The predation phase is carried out at this step. Initially, the radius of the hazard zone is determined utilizing Eq. (19). To improve the AI capacity to reliably identify voice emotion and more effectively enhance the user experience.

$$Q = \frac{\Sigma_{i=1}^{m}}{|w_j^{ka} - w_i^{va}|} \tag{19}$$

Here, m denotes the number of dimensions while w_j^{ka} and w_i^{va} stand is for the lower and upper bounds of the members, respectively. Following the radius's computation, a group of intended prey is identified as

$$S_{oj} = g_i \epsilon\, G | TU_{gj}, < TU_{oj}, | \left| o_j - g_i \right| | \leq Q, g_i \nexists \mathrm{L} \tag{20}$$

where $| \left| o_j - g_i \right| |$ indicates the Euclidean distance between the individuals o_j and g_j, and TU_{gj} and TU_{oj} indicate the survival values of pi and hj, respectively. Currently, each member of the set described in Eq. (10) has their probability of being hunted estimated, and set K is defined so that $\{L = L, g_i\}$.

$$G_{oj}, g_i = \frac{\omega_{o_j}, g_i}{\sum_{(g_n \in S_{o_j})} \omega_{o_j}, g_n}, g_i \in S_{oj} \tag{21}$$

where the prey attraction between o_i and g_i is represented by ω_{o_j}, g_i.

8. Creating a set M such that $N = g_i \in L$, where L is the set of herd members killed during the predation phase, is the final stage in the restoration process. For every member in N, corresponding matching probabilities are computed using.

$$O_q = \frac{TU_{g_j}}{\sum_{(g_n \in N)} TUg_n}, g_i \epsilon N \tag{22}$$

The last step involves using SHO's mating operation, $ormix([Gq1, Gq2, \ldots, Gqm])$, to substitute each $g_i \epsilon N$ with a new answer. Algorithm 1 shows the process of SHO-LSTM.

Algorithm 1: The process of SHO-LSTM

```
lstm_model = initialize_lstm_model()
predicted_emotions = lstm_model.predict(vocal_emotion_data)
train_lstm_with_SHO(lstm_model, vocal_emotion_data, SHO_parameters)
population_size = N
dimension = D
max_iterations = T
lower_bound = lb
upper_bound = ub
predator_positions = random_initialize_population(population_size, dimension, lower_bound, upper_bound)
prey_positions = random_initialize_population(population_size, dimension, lower_bound, upper_bound)
fitness_predators = evaluate_fitness(predator_positions)
fitness_preys = evaluate_fitness(prey_positions)
for t in range(max_iterations):
prey_group_size = calculate_prey_group_size(population_size)
predators, preys = divide_population(predator_positions, prey_positions, prey_group_size)
For prey in prey:
```

(Continued)

Algorithm 1 (continued)

update_prey_position(prey, predators)
fitness_preys = evaluate_fitness(preys)
for the predator in predators:
update_predator_position(predator, preys)
fitness_predators = evaluate_fitness(predators)
best_position = update_global_best_position(predators, preys)
evaluate_model_performance(predicted_emotions, true_emotions)

Vocal emotions are dynamically conveyed using SHO-LSTM. This proposed approach orchestrates expressive communication by striking a balance between group cohesiveness and individual self-interest. While long-term memory preserves context and ensures cohesive emotional narratives, short-term memory is better at capturing transitory details. This combination of group harmony and self-serving optimization gives voice expressions depth and genuineness.

4 Result

In this study, Windows 11 and the Python platform are employed. It evaluates the performance of the SHO-LSTM with existing methods such as Cumulative Attribute-Weighted Graph Neural Network (CA-WGNN) [21], Support vector machines (SVM) [21], Gradient Boosting Model (GBM) [21], and Voting Classifier (Logistic Regression and Stochastic Gradient Descent) (VC (LR-SGD)), with metrics like precision, recall, accuracy, and F1-score are used for assessment.

4.1 Accuracy

The accuracy with which a system can precisely translate spoken words into text or actions is known as voice recognition accuracy in HCI. It is usually computed as a percentage based on the ratio of words that are successfully identified to all words that are spoken. Table 2 and Fig. 5 show the accuracy comparison of the proposed method. Compared to the prior approaches such as VC (LR-SGD) at 79%, CA-WGNN at 94%, GBM at 74%, and SVM at 76%, the proposed approach SHO-LSTM has achieved superior accuracy of 97%. Thus, it proves enhanced user experience of AI-enabled communication.

Table 2: Numerical results of accuracy

Methods	Accuracy (%)
VC (LR-SGD) [21]	79
CA-WGNN [21]	94
GBM [21]	74
SVM [21]	76
SHO-LSTM [Proposed]	97

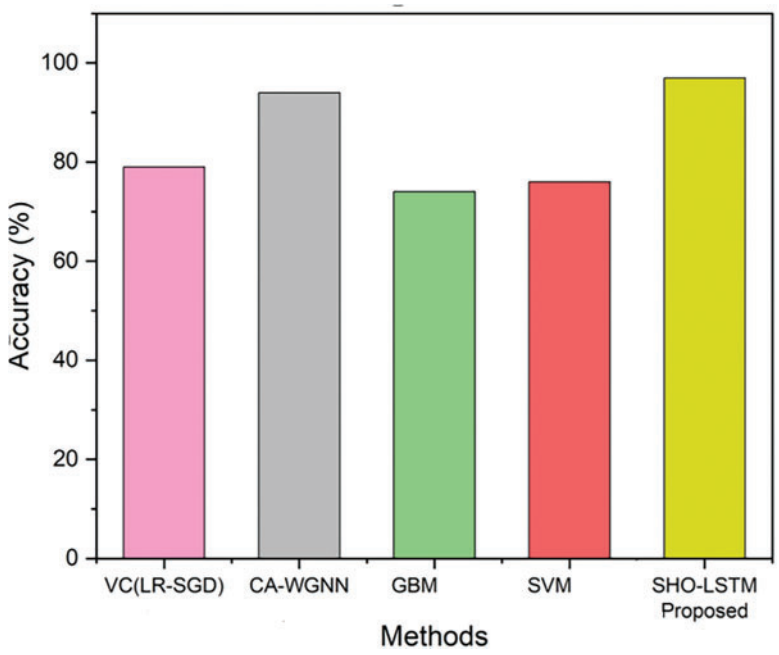

Figure 5: Comparison of accuracy

4.2 Precision

Efficiently measuring the precision of sentiment analysis algorithms yields the proportion of correctly recognized positive or negative sentiments among all anticipated positive or negative emotions. Maximizing accuracy can lead to more accurate and sensitive assessments of user input as well as emotions by enhancing the ability of the model to identify certain emotions. The precision comparison of the suggested approach is shown in Fig. 6 and Table 3. When comparing other existing methods such as SVM (76%), GBM (72%), CA-WGNN (92%), and VC (LR-SGD) (78%), the proposed method SHO-LSTM approach demonstrates higher accuracy with a 95% precision in identifying voice emotions.

4.3 Recall

To demonstrate the significance of recall as a critical performance metric for user sentiment prediction, assess the model's capacity to detect each occurrence of a positive or negative feeling in the dataset. Maximizing memory and accurately representing user emotions and input are crucial, as it is improving the model to locate as many instances of the goal mood as it is practically possible. Table 4 and Fig. 7 show the recall comparison of the proposed method. The suggested SHO-LSTM demonstrates the efficacy in improving emotion identification for better user experience in AI communication by achieving 96% recall, surpassing the other existing methods such as SVM (80%), GBM (79%), VC (LR-SGD) (84%), and CA-WGNN reached (93%).

4.4 F1-Score

The F1-score offers a thorough assessment of the model's capacity to recognize and categorize users' vocal emotions while accounting for accuracy and recall. Maximizing the F1-score leads to a comprehensive and dependable analysis of user feelings and remarks, but it requires striking

a compromise between exact sentiment identification coupled with extensive sentiment coverage. Table 5 and Fig. 8 show the accuracy comparison of the proposed method. With a 95% F1-score, the suggested SHO-LSTM approach outperformed the existing techniques such as VC (LR-SGD) (81%), SVM (78%), GBM (76%), and CA-WGNN (91%), ensuring improved voice emotion recognition for improved AI-human interaction.

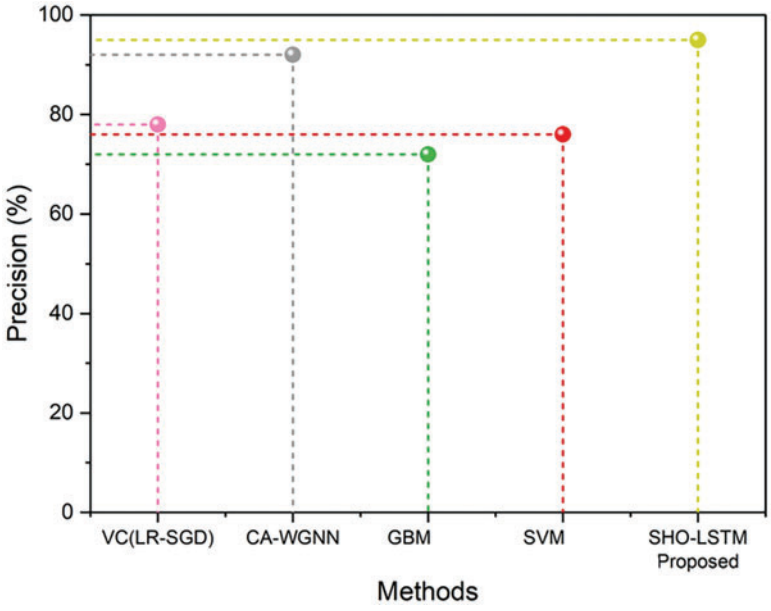

Figure 6: Comparison of precision

Table 3: Numerical results of precision

Methods	Precision (%)
VC (LR-SGD) [21]	78
CA-WGNN [21]	92
GBM [21]	72
SVM [21]	76
SHO-LSTM [Proposed]	95

Table 4: Numerical results of recall

Methods	Recall (%)
VC (LR-SGD) [21]	84
CA-WGNN [21]	93
GBM [21]	79
SVM [21]	80
SHO-LSTM [Proposed]	96

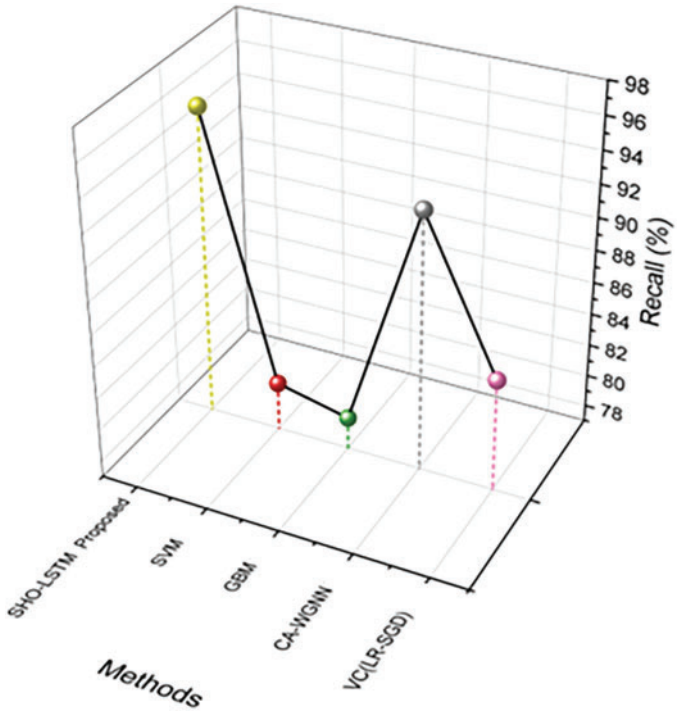

Figure 7: Comparison of recall

Table 5: Numerical results of F1-score

Methods	F1-score (%)
VC (LR-SGD) [21]	81
CA-WGNN [21]	91
GBM [21]	76
SVM [21]	78
SHO-LSTM [Proposed]	95

The accuracy and precision percentages of vocal emotion identification from using SHO-LSTM for different emotional states are displayed in Fig. 9. The proposed SHO-LSTM is particularly good at recognizing happiness, with an accuracy rate of 98%; neutral emotions come in second at 92%. But when it comes to identifying emotions like fear (85%), sorrow (78%), and anger (80%), it performs a little less well. Despite these fluctuations, most emotions have consistently high accuracy rates; surprise and tranquility have rates as high as 94% and 89%, respectively.

A confusion matrix compares actual data with prediction to assess how well a classification algorithm performed. Exhibiting true positives, true negatives, false positives, and false negatives, demonstrates how well the proposed approach works in enhancing the user experience in AI-powered human-computer interaction through voice recognition of emotions utilizing the machine learning approach. The RAVDESS and CREMA-D dataset's confusion matrices are displayed in Fig. 10a,b.

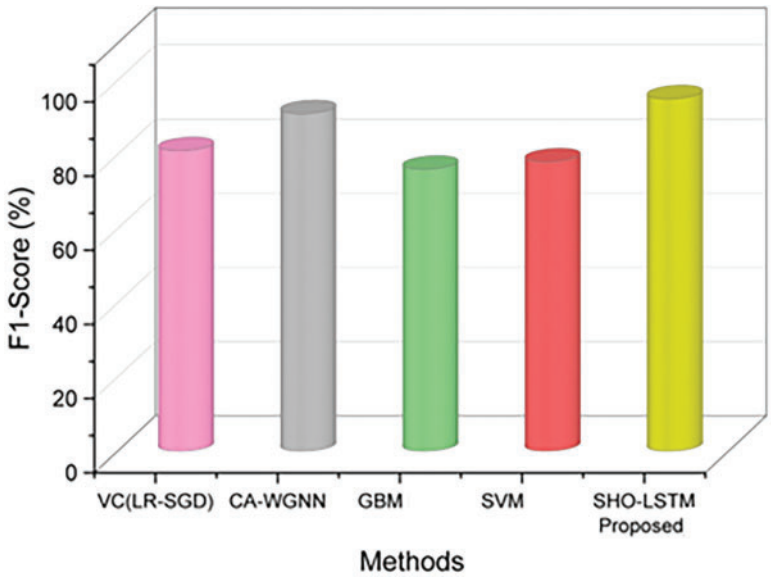

Figure 8: Comparison of F1-score

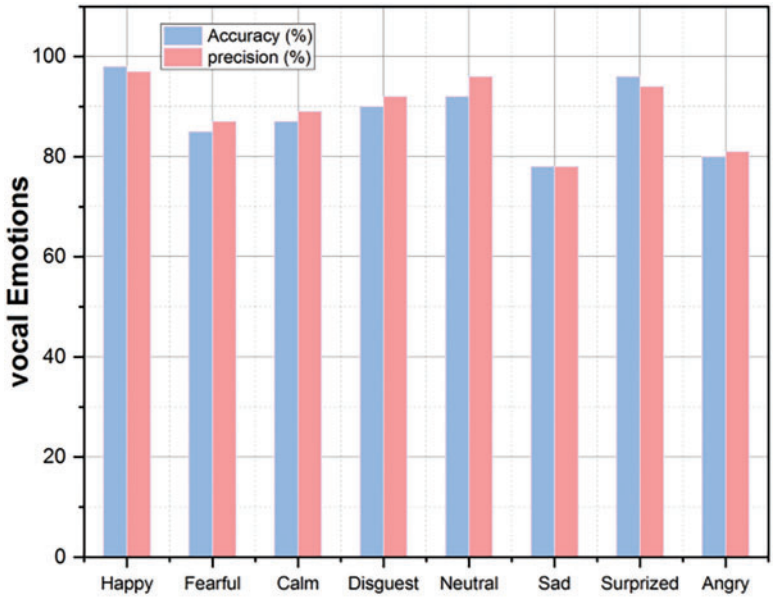

Figure 9: Vocal emotions predicted by proposed SHO-LSTM

The AUC (Area Under the Curve) methods indicate the strength of a classifier's approach to differentiate between classes. To improve user experience in AI-powered human-computer communication including Vocal Emotions Identification, a higher AUC indicates better performance of the novel deep learning method, SHO-LSTM, in accurately representing different vocal emotions, shown in Fig. 11.

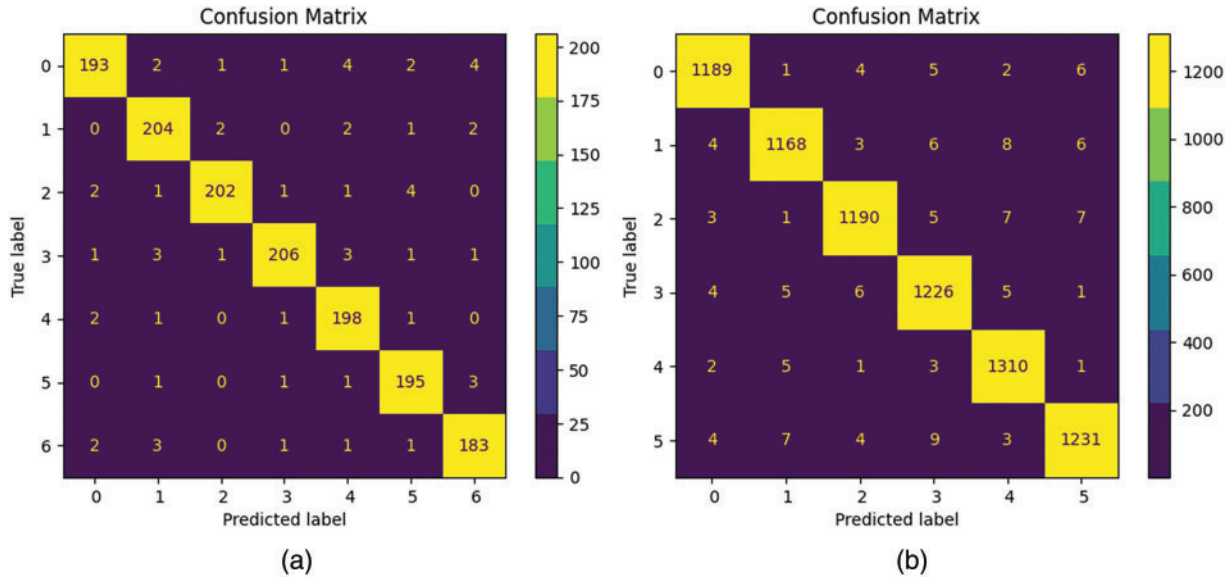

Figure 10: Confusion matrix comparison (a) RAVDESS and (b) CREMA-D

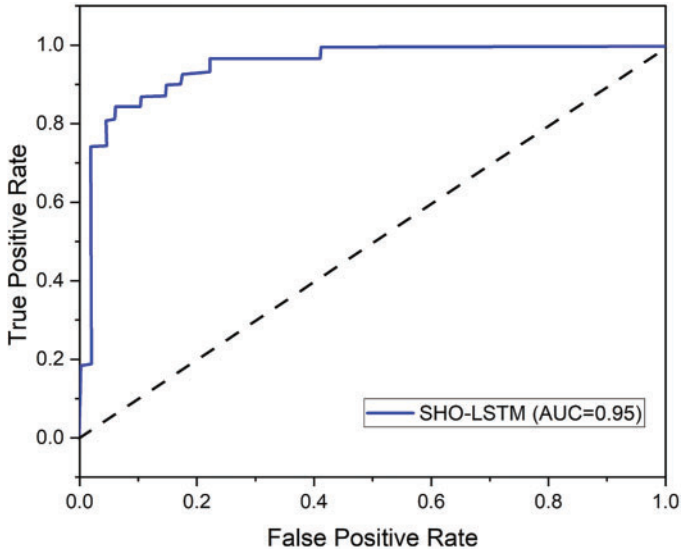

Figure 11: Outcome of AUC

4.5 Discussion

Voice modulation is a technological progression that enables computers to identify and express vocal emotions in HCI. This technology enhances user experience and assists with occupations like virtual assistants, therapy, and customer service by promoting sympathetic relationships. SVM's [21] inclination to overfit with high-dimensional data is one of its limitations when it comes to vocal emotion analysis. This might potentially result in worse generalization performance and make it more challenging to capture minute details in emotional expressions. GBM [21] can be computationally demanding and prone to overfitting, which might dampen the enthusiasm of those looking for quick

insights from their data despite its brilliance in predicting accuracy. By mixing different models, the Voting Classifier (LR-SGD) [21] can become less interpretable, which might obscure important information. Inconsistent predictions might result from this fusion if different models have conflicting biases. A limitation of the CA-WGNN [21] is its poor ability to encode and understand complex emotional signals due to its low capacity to capture subtle subtleties of vocal emotions. Using the SHO-LSTM technique resolves these problems. It overcomes limits in vocal emotion identification by proactively modifying model parameters, which also increases interpretability when compared to ensemble approaches and mitigates the overfitting tendencies of SVM and GBM. The RAVDESS dataset may not fully capture all forms of voices or emotions accurately, thereby making the model biased or unreliable in real-world circumstances. The proposed approach might function well in the present dataset, but it may struggle when applied to large or complicated datasets from real-life scenarios in which emotions are conveyed differently or more subtly. While the proposed deep learning approach improves the emotion recognition task in human-computer interaction, several limitations might inhibit the approach. Sensitivity to noise, such as background noise or overlapping speech, could undermine the actual accuracy of performance in a real-world scenario. Real-time performance is an accomplishment that is yet to be achieved due to the computational complexity involved in processing vocal cues. Addressing these issues is essential to ensuring robust and equitable user experiences.

5 Conclusion

Voice is a technological advance in human-computer interaction (HCI) that enables computers to perceive and mimic vocal emotions. This technology enhances user experience and assists in areas such as virtual assistants, therapy, and customer service by creating sympathetic relationships. It modifies communication dynamics by bridging the gap between humans and robots. This article uses a SHO-LSTM strategy to identify emotions in human communication. The RAVDESS dataset collection that will be obtained will be used in our suggested security identification technique. A Wiener filter will be used for image denoising quality enhancement and feature extraction using MFCC. To analyze the proposed strategy employing several measures, including precision (95%), accuracy (97%), recall (96%), and F1-score (95%), the recommended model's detection ability will be assessed during the finding assessment phase.

5.1 Limitation

The suggested method's capacity to generalize well across all contexts is constrained by its reliance on the RAVDESS dataset, which might not fully represent the spectrum of vocal emotions across many languages and cultures.

5.2 Future Scope

To improve the system's capability to recognize emotions, future studies should investigate using multimodal data, such as text, physiological signals, and facial expressions. Furthermore, using a variety of datasets from different contexts and languages, together with sophisticated noise mitigation techniques, might increase the model's generalizability and resilience.

Acknowledgement: The author Dr. Arshiya S. Ansari extends the appreciation to the Deanship of Postgraduate Studies and Scientific Research at Majmaah University for funding this research work through the project number (R-2025-1538).

Funding Statement: The author Dr. Arshiya S. Ansari extends the appreciation to the Deanship of Postgraduate Studies and Scientific Research at Majmaah University for funding this research work through the project number (R-2025-1538).

Author Contributions: All authors have equal contribution in this research. Data collection: Ahmed Alhussen, Mohammad Sajid Mohammadi, Arshiya Sajid Ansari; Methedoogy: Arshiya Sajid Ansari, Ahmed Alhussen, Mohammad Sajid Mohammadi; Draft manuscript preparation: Arshiya Sajid Ansari, Mohammad Sajid Mohammadi, Ahmed Alhussen. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable. All references are from Google Scholar.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

[1] A. A. Alnuaim *et al.*, "Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier," *J. Healthc. Eng.*, vol. 2022, no. 2, pp. 1–12, 2022. doi: 10.1155/2022/6005446.

[2] P. Chhikara, P. Singh, R. Tekchandani, N. Kumar, and M. Guizani, "Federated learning meets human emotions: A decentralized framework for HCI for IoT applications," *IEEE Internet Things J.*, vol. 8, no. 8, pp. 6949–6962, 2020. doi: 10.1109/JIOT.2020.3037207.

[3] Z. Lv, F. Poiesi, Q. Dong, J. Lloret, and H. Song, "Deep learning for intelligent HCI," *Appl. Sci.*, vol. 12, no. 22, 2022, Art. no. 11457. doi: 10.3390/app122211457.

[4] S. B. Velagaleti, D. Choukaier, R. Nuthakki, V. Lamba, V. Sharma and S. Rahul, "Empathetic algorithms: The role of AI in understanding and enhancing human emotional intelligence," *J. Electr. Syst.*, vol. 20, no. 3s, pp. 2051–2060, 2024. doi: 10.52783/jes.1806.

[5] K. Loveys, M. Sagar, and E. Broadbent, "The effect of multimodal emotional expression on responses to a digital human during a self-disclosure conversation: A computational analysis of user language," *J. Med. Syst.*, vol. 44, no. 9, 2020, Art. no. 143. doi: 10.1007/s10916-020-01624-4.

[6] Y. Wang, "Research on the construction of human-computer interaction system based on a machine learning algorithm," *J. Sens.*, vol. 2022, no. 2, pp. 1–11, 2022. doi: 10.1155/2022/3817226.

[7] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. Sadeeq, and S. Zeebaree, "Multimodal emotion recognition using deep learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 1, pp. 73–79, 2021. doi: 10.38094/jastt20291.

[8] T. Numata, Y. Asa, T. Hashimoto, and K. Karasawa, "Young and old persons' subjective feelings when facing with a non-human computer-graphics-based agent's emotional responses in consideration of differences in emotion perception," *Front. Comput. Sci.*, vol. 6, 2024, Art. no. 1321977. doi: 10.3389/fcomp.2024.1321977.

[9] Z. Yang, X. Jing, A. Triantafyllopoulos, M. Song, I. Aslan and B. W. Schuller, "An overview & analysis of sequence-to-sequence emotional voice conversion," 2022. doi: 10.48550/arXiv.2203.15873.

[10] F. Ferrada and L. M. Camarinha-Matos, "Emotions in human-AI collaboration," in *Navigating Unpredictability: Collaborative Networks in Non-Linear Worlds*. Cham, Springer Nature Switzerland, 2024, pp. 101–117. doi: 10.1007/978-3-031-71739-0_7.

[11] J. J. Sundi, H. Kumar, and R. Bedi, "Real-time facial expression recognition using convolutional neural networks for adaptive user interfaces," in *2024 5th Int. Conf. Emerg. Technol. (INCET)*, IEEE, 2024, pp. 1–6. doi: 10.1109/INCET61516.2024.10593062.

[12] J. Suo *et al.*, "Enabling natural human-computer interaction through AI-powered nanocomposite IoT throat vibration sensor," *IEEE Internet Things J.*, vol. 11, no. 14, pp. 24761–24774, 2024. doi: 10.1109/JIOT.2024.3382101.

[13] W. Alsabhan, "HCI with real-time SER with ensembling techniques 1D convolution neural network and attention," *Sensors*, vol. 23, no. 3, 2023, Art. no. 1386. doi: 10.3390/s23031386.
[14] Y. Yoshitomi, "Human-computer communication using recognition and synthesis of facial expression," *J. Robot. Netw. Artif. Life.*, vol. 8, no. 1, pp. 10–13, 2021. doi: 10.2991/jrnal.k.210521.003.
[15] X. Chen, L. Xu, M. Cao, T. Zhang, Z. Shang and L. Zhang, "Design and implementation of human-computer interaction systems based on transfer support vector machine and EEG signal for depression patients' emotion recognition," *J. Med. Imaging Health Informat.*, vol. 11, no. 3, pp. 948–954, 2021. doi: 10.1166/jmihi.2021.3340.
[16] Y. Du, R. G. Crespo, and O. S. Martínez, "Human emotion recognition for enhanced performance evaluation in e-learning," *Prog. Artif. Intell.*, vol. 12, no. 2, pp. 199–211, 2023. doi: 10.1007/s13748-022-00278-2.
[17] C. Jie, "SER based on convolutional neural network," in *2021 Int. Conf. Netw., Commun. Inf. Technol. (NetCIT)*, IEEE, Dec. 2021, pp. 106–109.
[18] Y. Du, K. Zhang, and G. Trovato, "Composite emotion recognition and feedback of social assistive robot for elderly people," in *Int. Conf. Human-Comput. Interact.*, 2023, pp. 220–231. doi: 10.1007/978-3-031-35894-4_16.
[19] N. Gasteiger, J. Lim, M. Hellou, B. A. MacDonald, and H. S. Ahn, "A scoping review of the literature on prosodic elements related to emotional speech in human-robot interaction," *Int. J. Soc. Robot.*, vol. 16, no. 4, pp. 1–12, 2022. doi: 10.1007/s12369-022-00913-x.
[20] T. Han, Z. Zhang, M. Ren, C. Dong, X. Jiang and Q. Zhuang, "SER based on deep residual shrinkage network," *Electronics*, vol. 12, no. 11, 2023, Art. no. 2512. doi: 10.3390/electronics12112512.
[21] H. F. T. Al-Saadawi and R. Das, "TER-CA-WGNN: Trimodel emotion recognition using cumulative attribute-weighted graph neural network," *Appl. Sci.*, vol. 14, no. 6, 2024, Art. no. 2252. doi: 10.3390/app14062252.