Enhanced Emotion Recognition Using a Hybrid Autoencoder-LSTM Model Optimized with a Hybrid ACO-WOA Algorithm for Hyperparameter Tuning

Vinod Waiker[1], Janjhyam Venkata Naga Ramesh[2], Ms Kiran Bala[3],
Dr. V.V. Jaya Rama Krishnaiah[4], Dr.T. Jackulin[5], Elangovan Muniyandy[6], Osama R.Shahin[7]
Datta Meghe Institute of Management Studies, Nagpur, Maharashtra, India[1]
Adjunct Professor, Department of CSE, Graphic Era Hill University, Dehradun, 248002, India[2]
Adjunct Professor, Department of CSE, Graphic Era Deemed To Be University, Dehradun, 248002, Uttarakhand, India[2]
Lecturer, Department of Computer Science-College of Engineering and Computer Science, Jazan University, Jazan, KSA[3]
Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India[4]
Associate Professor, Department of CSE, Panimalar Engineering College, Chennai, Tamil Nadu, India[5]
Department of Biosciences-Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai - 602 105, India[6]
Applied Science Research Center, Applied Science Private University, Amman, Jordan[6]
Department of Computer Science-College of Computer and Information Sciences, Jouf University, Saudi Arabia[7]
Physics and Mathematics Department-Faculty of Engineering, Helwan University, Helwan, Egypt[7]

***Abstract*—Emotion recognition is vital in the human Computer interaction because it improves interaction. Therefore, this paper proposes an improved method for emotion recognition regarding the Hybrid Autoencoder-Long Short-Term Memory (LSTM) model and the newly developed hybrid approach of the Ant Colony Optimization (ACO) and Whale Optimization Algorithm (WOA) for hyperparameters tuning. In this case, Autoencoder can reduce input data dimensionality for input data and find the features relevant for the model's work. In addition, LSTM is able to work with temporal structures of sequential inputs like speech and videos. The contribution of this research lies in the novel combination method of ACO-WOA which aims at tweaking hyperparameters of Autoencoder-LSTM model. Global aspect of ACO and WOA thereby improve the search efficiency and the accuracy of the proposed emotion recognition system and its generalization capacity. In context with the benchmark dataset for the experimentations of emotion recognition, it has established the efficiency of the proposed model in terms of the conventional methods. Recall rates in recognitive intended various emotions and different modalities were also higher in the hybrid Autoencoder-LSTM model. The optimization algorithms like the ACO-WOA also supported in reducing the computational cost which arose due to hyperparameters tuning. The implementation of this paper is done through Python Software. This implementation shows a high accuracy of 94.12% and 95.94% for audio datasets and image datasets respectively when compared with other deep learning models of Conv LSTM and VGG16. Therefore, the research shows that the presented hybrid approach can be a useful solution for successfully employing emotion recognition for enhancing the creation of the empathetic AI systems and for improving user interactions within various fields including healthcare, entertainment, and customer support.**

Keywords*—Emotion recognition; autoencoder; long short-term memory; Ant Colony Optimization (ACO); Whale Optimization Algorithm (WOA)

I. INTRODUCTION

The development of machine learning methods has made it simple for computers to recognize and comprehend how humans act using a variety of approaches. One of the essential components of human conduct is emotions [1]. A vast range of programs, including political analysis, advertising, and interactions between humans and computers, are improved by the identification of individual feelings or emotions. People now share feelings and knowledge on online social networks (OSNs) like Facebook, Instagram, Twitter, and other online social networks as part of their everyday routines [2]. These OSNs' abundance of data and information makes them ideal for researching and analyzing human emotions and behavior. It also encourages the development of more emotion-aware apps. A customized system for recommendations suggests tailored items; it suggests videos, music, or movies based on the user's preferences and feelings [3]. In times of disaster or epidemic, recognition of emotions is used to assess public opinion. This information aids in decision-making and situational management on the part of the government. Because of this, ONSs now use emotion detection as a distracting and finding faults activity. A person's motivation, state of emotions, psychological disorders, and level of mental activity may all be inferred from their facial expressions [4].

The facial expressions are a powerful expression and communication tool in interpersonal relationships[5]. The ability of Facial Emotion Recognition (FER) to characterize an individual's feelings or psychological state lends credence to its significance [6]. Its uses go beyond only analyzing human behavior, assessing someone's emotional condition, or assessing someone's psychological wellness [7]. Additionally, it is making inroads into a variety of other industries, including automation, schooling, holography, intelligent medical systems, safety

systems, law enforcement, amusement, multimodal interaction and stress identification [8]. The inclusion of movements of the face in these domains demonstrates the significance of facial emotions in human existence. These days, one of the hardest problems in computational science is automated FER [9]. Movements and spoken words can be used to communicate emotions. It is not only dependent on facial features. Just 7% of the background of the data may be conveyed verbally; the remaining 38% can be conveyed by voice tone, cadence, and speaking rate [10]. Conversely, around 55% of information is conveyed by facial expressions. A person's facial expressions may reveal a lot about their mental health. Facial expressions are used in many facets of life and are not only restricted to certain professions [11]. In the field of health disciplines, bipolar patients benefit from FER. Physicians are attempting to identify and track the behavior of their clients, including the feelings and actions of bipolar patients throughout their illness [12].

Many sophisticated FER methods have been developed that allow the system to recognize human facial expressions when it receives facial pictures as input [13]. Humans may express themselves in seven different ways: fear, happiness, surprise, neutral, anger, sadness, and contempt [14]. Multifunctional emotional line databases have been used in this work to increase the classification accuracy of emotions. To achieve the ultimate goal, three main components—preprocessing, extraction of features, and classification—are further subdivided. The selected information in this procedure includes pictures, sounds, and videos that were taken from a variety of individuals, comprising men and women. Every picture is taken from the front and is separated into eight groups [15]. To ensure that every image is the same size, the initial step of preliminary processing involves reshaping each image to measure 150 × 150 pixels [16]. Additionally, photos are automatically enlarged and flipped among 0- and 180-degrees during preprocessing. Moreover, pictures are rotated both vertically and horizontally. Pictures are further analyzed to obtain attributes in the next stage. Next, important variables that are essential for the algorithm's applicability are retrieved from all of the recorded video and audio data. Only valuable features should be retained once the features have been retrieved, and max pooling aids in this process [17]. The final stage of the suggested process, categorization, is in charge of identifying the accurate labels. completely linked layers are employed in categorization, and these entirely interconnected layers additionally use two layers that are hidden. There are many weighted nodes in each hidden layer [18]. The weight value of each node increases with bias values through forward propagation procedures before the total calculation is performed. The algorithm performs backpropagation to identify the actual label of an input picture through adjustments to the hidden layer node weights [19]. The following sections include the key contribution of the paper.

- The research introduces a novel hybrid model that combines Autoencoders and LSTM networks to enhance emotion recognition. The Autoencoder effectively reduces dimensionality and extracts salient features, while the LSTM component captures temporal dependencies in sequential data, providing a comprehensive approach to emotion analysis.
- A significant contribution is the creation of a hybrid optimization algorithm that merges ACO and WOA for hyperparameter tuning. This hybrid approach balances global and local search capabilities, leading to more efficient and accurate identification of optimal hyperparameters for the model.
- The proposed model demonstrates superior performance in recognizing emotions across various modalities, including speech, facial expressions, and physiological signals. This improvement in accuracy is attributed to the effective combination of the Autoencoder-LSTM model and the optimized hyperparameters found through the hybrid ACO-WOA algorithm.
- The research highlights a reduction in the computational cost associated with hyperparameter tuning. The hybrid ACO-WOA algorithm accelerates the optimization process, making it feasible to apply deep learning models to emotion recognition tasks without the typically prohibitive computational overhead.
- The findings suggest broad applicability of the enhanced emotion recognition system in fields such as healthcare, education, customer service, and entertainment. The research provides a foundation for developing more empathetic and responsive AI systems, capable of understanding and interacting with users based on their emotional states.

The paper continues with its structure by explaining Section II. Section III describes the problem statement which will be addressed by the proposed paper. Section IV detailed the construction process of Autoencoder-LSTM network. A performance evaluation of the proposed Autoencoder-LSTM network takes place in Section V before the article's conclusion in Section VI.

II. Related Works

The identification and treatment of certain medical diseases may alter if neurological signals are used to recognize emotions [20]. Generalized emotional detection programs may have issues and limits because of the limited amount of facial movement factors persons who fake their feelings, or those who have alexithymia. By examining the constant neurons produced by the human brain, these signals may be found. Brainwaves known as EEGs provide with a more comprehensive understanding of the psychological emotions that people might be unable to articulate. Neuronal communication channels can cause modifications to electrical potential, which might be reflected in brainwave EEG data. This study compares several artificial intelligence approaches, including SVM, K-nearest neighbour, Linear Discriminant Analysis, LR, and DT. Each of these algorithms is evaluated using and without principal component analysis for reducing the dimensionality. The historic information gathered from EEG sensor networks is analyzed. In order to reduce the duration of execution, grid computing was also used for hyper-parameter tweaking for each of the models created using machine learning that were evaluated over Spark cluster. This investigation made use of the multidimensional DEAP Information set, that is designed for the examination of individual emotional states.

The paperwork seeks to generate an artificial intelligence structure for programmed emotion recognition from words [21]. The established structure is to be utilized in the structure of tracking public sentiments. A short evaluation of additional investigation articles on the procedure of establishing artificial intelligence frameworks for effortless sentiment recognition from conversation has been specified in the document. Traditional and deep machine learning approaches and techniques and certain characteristics of the original information set have been taken into account. The DailyDialog and its effectiveness for training the classificatory have been considered. Furthermore, constructing and identifying the ideal framework for natural sentiment recognition from conversation has been suggested. The study's findings on the effects of variables like the quantity of documents in every group in the training set of data, content pre-processing, vectorization or word-integration techniques, artificial intelligence approach selection for identifying text, parameter settings, and structure are provided. The previous section provided demonstrations of how to use the artificial intelligence algorithm to analyze the actual information that was gathered. It has been demonstrated how certain occurrences in the lives of society, the people living in a specific region, or a community are correlated with changes in the quantity of data falling into various psychological classifications. Lastly, the artificial intelligence algorithm's shortcomings and a few potential improvements to the framework for identifying emotions have been discussed.

Accurate emotion detection from speech signals contributes to improved HCI [22]. The extracted characteristics from language signals determine how well a SER algorithm performs. But since the efficacy of characteristics vary with feelings, choosing the best collection of depictions of features in SER continues to be the most difficult challenge. The worldwide long-term situational descriptions of language signals are ignored in many investigations that identify concealed specific language aspects. Due to inadequate representations of attributes and a lack of readily accessible information the current SER method performs poorly in detection tasks. Inspired by CNN, LSTM, and GRU's effective extracting features, this paper suggests a combination that makes use of the overall predictive capabilities of three distinct designs. Initially 1D CNN is used in the design, and then FCN are used. The CNN network is followed by the LSTM-FCN and GRU-FCN layers in the other two designs, accordingly. The goal of each of the three distinct frameworks is to derive voice waves' local and over time worldwide situational expressions. The weighted mean of the various models is used by the ensembles. In order to improve system generalizations, the information in this work have been enhanced by adding additional white Gaussian noise, pitch shifting, and noise level stretching.

Accurate emotion detection from speech signals contributes to improved HCI [22]. The extracted characteristics from language signals determine how well a SER algorithm performs. But since the efficacy of characteristics vary with feelings, choosing the best collection of depictions of features in SER continues to be the most difficult challenge. The worldwide long-term situational descriptions of language signals are ignored in many investigations that identify concealed specific language aspects. Due to inadequate representations of attributes and a lack of readily accessible information the current SER method performs poorly in detection tasks. Inspired by CNN, LSTM, and GRU's effective extracting features, this paper suggests a combination that makes use of the overall predictive capabilities of three distinct designs. Initially 1D CNN is used in the design, and then FCN are used. The CNN network is followed by the LSTM-FCN and GRU-FCN layers in the other two designs, accordingly. The goal of each of the three distinct frameworks is to derive voice waves' local and over time worldwide situational expressions. The weighted mean of the various models is used by the ensembles. In order to improve system generalizations, the information in this work have been enhanced by adding additional white Gaussian noise, pitch shifting, and noise level stretching.

One of the most important things that can reveal an individual's emotional state is their facial expression [23]. Individuals can communicate vocally for around 45% of the time, and nonverbally for about 55% of the time. One of the hardest problems in technology right now is automated face expressions detection. FER has several uses outside of analyzing behaviour and keeping tabs on people's emotions and psychological wellness. It is also making inroads into other domains, including learning, robotics, entertainment, holography, smart medical systems, safety technologies, criminal justice theory, and identifying stress. Emotions on the face are becoming increasingly significant in medical studies, especially for bipolar patients whose mood fluctuations are common. This paper suggests a computerized structure and algorithms for facial recognition utilizing a CNN that has two layers that are hidden and four convolution layers for enhanced precision. Various face pictures of males and females with emotions including rage, anxiety, resentment, dislike, neutral, joyful, sorrowful, and surprised are included in an expanded collection. Three main processes are included in this research's implementation of FD-CNN: preprocessing, feature extraction, and classification. With this suggested approach, a 94% FER precision is attained. K-fold cross-validation is used to verify the suggested approach.

These days, neural networks, deep learning, and algorithmic learning are the main tools used to enhance a device's ability [8]. The intelligent SER algorithm is a fundamental requirement and a developing field of study in digital voice analyzing; yet, SER performs a significant role with numerous purposes associated with HCI. In order to make the most advanced SER structure workable for actual time business purposes, it must be improved. The main cause of inadequate precision and poor forecasting rate is the scarcity of information and an algorithm arrangement that is the hardest part of trying to create a strong machine learning approach. The constraints of the current SER methods were discussed in this research, and suggested a novel AI-based system design for the SER that makes use of the structural modules of the ConvLSTM with sequential learning. The local features learning block (LFLB), one of the four ConvLSTM blocks created in this study, and is used for obtaining regional psychological traits in a hierarchy association. Convolution processes are used to derive visual signals, and the ConvLSTM layers are chosen for input-to-state and states to states transitions. Utilizing the residual learning technique, this work

deployed four LFLBs to derive the spatiotemporal cues from the hierarchy correlated type voice signals.

The papers discuss diverse approaches for recognising emotions with the help of artificial intelligence based on neurological signals, voice, and face. Some of the issues that are in disagreement with the general emotion recognition are the so-called fake emotions, or "alexithymia'', and the use of the EEG brainwaves for more accurate recognition of the emotions. The basic BCI MLS algorithms are SVM, KNN, LDA and Decision Trees with and without applying PCA on EEG data of ALS subjects. Another work discusses the automated emotion recognition from speech which employs CNN, LSTM, GRU frameworks and discusses the challenges involved in feature selection. FER based on CNNs is introduced with high accuracy for emotions identification with focus on mental health assessment. Lastly, about the limitations encountered in the current SERs that are based on speech, a new SER approach that employs ConvLSTM layers is also presented for enhanced real-time performance. Research reviews that point out feature extraction as well as model optimization as crucial areas to enhance AI-based emotion detection form the basis of all the studies.

III. Problem Statement

Due to the temporal structure of emotional data and the complexity and diversity of emotional displays, emotion identification algorithms have difficulty correctly detecting emotional states. In many conventional approaches, handling increased dimensionality of input data and proper tuning of hyperparameters to determine performance and robustness of the Emotion recognition system remains a problem [8]. This research, therefore seeks to develop a solution by developing an emotion recognition model by combining Autoencoder and LSTM networks. For efficient feature extraction the Autoencoder is utilized whereas the LSTM component to enable temporal analysis of the features extracted from the emotional data. In order to increase the efficiency of the proposed model, hyperparameters tuning is performed using Ant Colony Optimization and Whale Optimization Algorithms. By combining LSTM for modeling temporal dependency and Autoencoder for dimensionality reduction, the proposed method enhances emotion recognition. LSTM incorporates the sequential nature of emotions, which reinforces the ability of the model to recognize complex emotional patterns over time, and the Autoencoder provides the benefits of noise removal and extraction of relevant features. In addition, the hyperparameters of the model are optimized well by the Hybrid ACO-WOA algorithm. It ensures improved accuracy and faster convergence without the threat of local minima through the strengths of Ant Colony Optimization and Whale Optimization. An improved accurate, efficient, and computationally efficient emotion recognition model is the result of this synergy.

IV. Proposed Autoencoder-LSTM Framework for Emotion Recognition

The research aims of furthering the capabilities of emotion recognition technology using a complex machine learning strategy. The work presents the novel Autoencoder-LSTM model for energy load prediction. The Autoencoder effectively learns and condenses salient characteristics from large complicated emotional data and the LSTM deciphers them with an understanding of time sequences, subtle emotional trends. Thus, in the present study, a combination of ACO and WOA is used in the form of hybrid optimization approach for better model performance and tuning of hyperparameters. Thus, the described strategy is based on the further enhancement of the model's parameters, providing increased accuracy of the emotions classification as well as providing the robustness of the model. In an attempt to increase the accuracy of the proposed models and address issues with hyperparameter tuning the research aims at improving emotion recognition in relation to human-computer interaction and potential use cases in mental health. Block Diagram for Autoencoder-LSTM is depicted in Fig. 1.

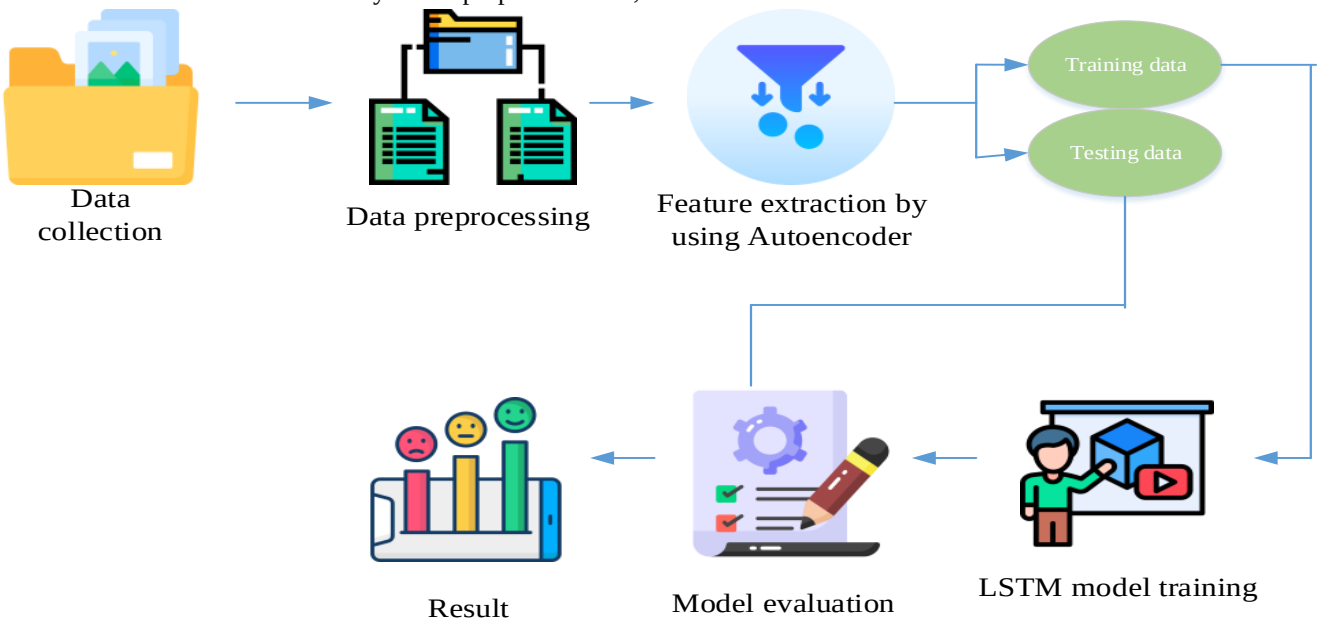

Fig. 1. Block diagram for autoencoder-LSTM.

A. Data Collection

The researchers upgraded and developed EmotionLines database into MELD by including textual information as well as audio-visual content which matches the original conversations. MELD hosts 1400 dialogues and 13,000 lines of dialogue originated from the Friends television show. Multiple speakers engaged in the recorded dialogues. Each statement within the discussions received assignment from among the seven recognized emotions including Anger, Resentment, Anxiety, Happiness, Neutral, Surprise and Fear. MELD provides emotion tags for its statements using three possible classifications: good, poor or unbiased [24] [25].

B. Data Pre-processing

1) Data cleaning: The most important step essential for the preparation of the data for deep learning is data cleaning which includes the detection of the errors and the removal of faults, contradictions, and mistakes in the datasets. This also requires that in order to feed the model with similar inputs, pixel values are normalized to a standard scale and noise is filtered out of the pictures. Due to this, the accuracy and reliability of raw data that is usually erroneous, and full of discrepancies, data cleaning becomes inevitable.

2) Data normalization: Normalization in image data preprocessing is the procedure which alters the direction and distribution of pixel values. To do that this step is conducted in order to match images from different sources and enhance the performance of machine learning models. For instance, min-max normalization rescales pixel values to a given range of 0 to 1 or -1 to +1 while z-score normalization assigns a pixel value based on the mean and standard deviation, and then transform it to standard normal distribution. Normalization enables a reduction of the impact of variety lighting condition, sensors and imaging system thus making images more comparable This is critical in research domain where due to variation in the imaging devices and techniques, differences in quality can make a lot of difference. Normalization if done by standardizing pixel values enhances the capability of image analysis, and machine learning models, and enhance the chances of deriving accurate results for tasks such as emotion recognition. Normalization can be mathematically expressed as in Eq. (1).

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

3) Noise reduction by weiner filter: For the noise removing part, this paper utilizes the Weiner Filter. The Statistical filtering issue has an optimum approach that is Wiener filter Filtration mechanism. Like mentioned before, this paper aims at explaining how statistical approach helps in solving the problem of linear filtering. At the same time, this paper agrees that noise has to be rejected and that properties of the target signal have to be assessed. Thus, it needs to be built in a way that the noise of the data which is fed into the filter causes as less an impact to the filter as possible. As for the linear filtering problems, the corresponding strategy is the minimization of the mean square error signal, which is the difference between the desired signal, to be transmitted, and the filtering outcomes. It is expressed as in Eq. (2).

$$e_k = y_k - \sum_{i=0}^{N-1} w(i).x_{k-i} \quad (2)$$

C. Proposed Autoencoder-LSTM Framework for Emotion Recognition

1) Feature extraction by using autoencoder: An Autoencoder is made up of an output component called the decoder and an input component called the encoder. The quantity of neurons in the encoder and decoder is identical. In addition, in comparison to the layers that provide input and output, the Autoencoder has a minimum a single hidden layer and fewer neurons. An Autoencoder's fundamental assumption is that, at the outcome, it needs to be capable to rebuild its input source using the lower-dimensional latent encoding of the information provided in the hidden layer. By quantifying the transformation losses or inaccuracy among the real source and its reorganized results, finding anomalies uses Autoencoder. A representative autoencoder architecture is seen in Fig. 3

2) 6. For data X, the goal function is to identify weight transmitters for decoder and encoder to reduce the reconstructing loss. It is expressed as in Eq. (3), (4), (5), (6) and (7).

$$\emptyset = X \to h \quad (3)$$

$$\Psi = h \longrightarrow X' \quad (4)$$

$$h = \sigma(W_{x+b}) \quad (5)$$

$$\emptyset, \psi = \arg\min \| X - (\psi * \phi)X \|^2 \quad (6)$$

$$Anomaly\ score = f(|X' - X|) \quad (7)$$

Where, h denotes latent representation,σ denotes activation function. W denotes weight matrix and b denotes bias vector.

3) Classification by Using LSTM: The vanishing gradient issue with the fundamental RNNs was the reason behind the creation of LSTM. Each LSTM networking cell has an extended-duration memory module connected to the cell state. Special gates, such as inputs, outputs, and forget gates, can be used to alter the present condition of this cell. These mechanisms allow it to selectively retain and erase information once it has been collected. This mechanism determines which data needs to be stored and which ones should be deleted. One of the main benefits of the LSTM network is its capacity to comprehend the long-term reliance of a sequence of information. Because of this characteristic, LSTM networks are the most popular kind of neural network for a variety of programs, including time-lapse forecasting, recognition of speech, processing of natural languages, and the ability to take in information consecutively.

Determining which aspects of the cell state need to be kept and which should be eliminated is the primary duty of the forget gate. This makes it easier for the LSTM framework to examine more closely, recognize when the form of the waves has changed

significantly, perceive more information, and keep an eye on extraneous data. Areas involving the ailments, including time series prediction, n-gram models of languages, and speech recognition, would benefit from this. When the score is near 1, it indicates that the details must be kept, and when it is around 0, it indicates that the data needs to be eliminated. Eq. (3) gives the mathematical formula for the forget gate. It is expressed as in Eq. (8).

$$f_t = \sigma\left(W_f * [h_t - 1, x_t] + b_f\right) \quad (8)$$

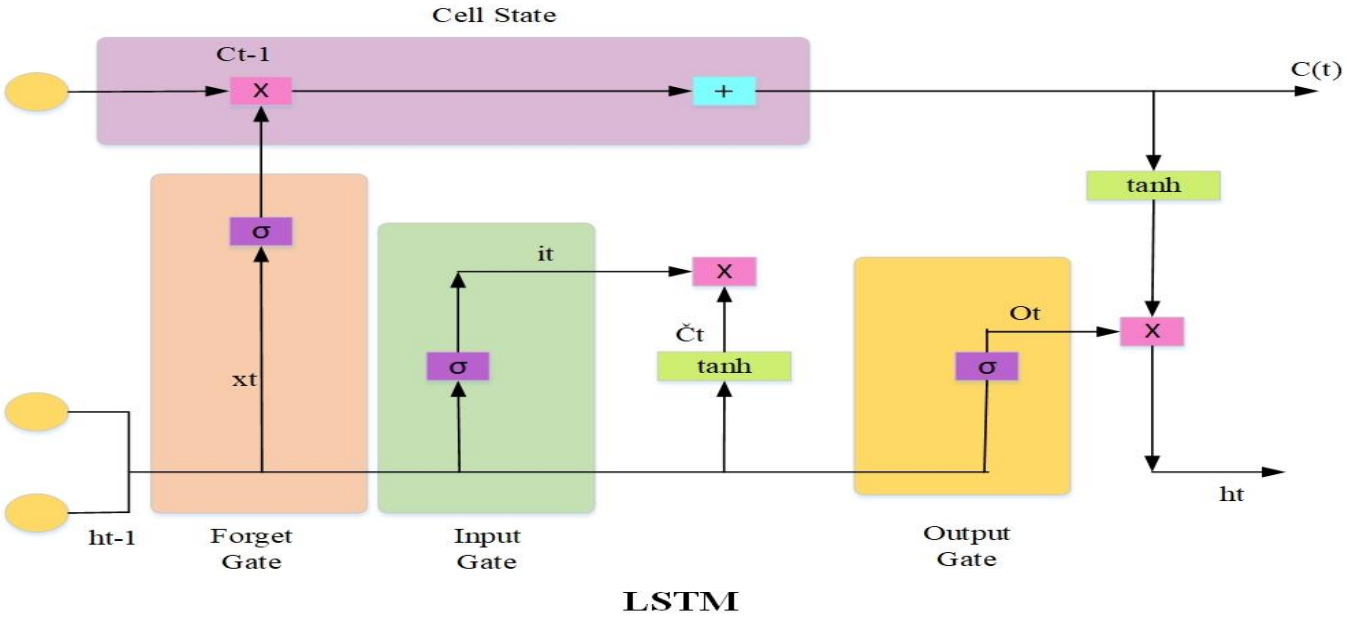

Fig. 2. LSTM Architecture diagram.

Fig. 2 represents the architecture diagram for the LSTM network. The input gate, in addition to the forget gate, determines what additional data has to be entered into the cell state. Candidate cell state and gate activation are its two primary constituents. This study demonstrates that the input gate's primary role is to control additional information's self-access at various cell states, allowing the LSTM system to create and develop novel components. By continuously altering the cell states, the input gate safeguards the long-term issues in the LSTM and assists it in retaining the firsthand info gathered from input series when needed. Such LSTM models may be used to train and modify activities involving extended-term contextual knowledge, including voice recognition, translation by machines, and time-series prediction. The mathematical equation for the input gate is expressed as in Eq. (9).

$$i_t = \sigma(W_i * [h_t\text{-}1, x_t\ ] + b_i\ ) \quad (9)$$

Eq. (10) computes the quantities to be introduced to the cell state that are valuable to the candidate's cell.

$$\hat{C}_t = tanh(W_c[h_{t-1}, x_t] + b_c) \quad (10)$$

The input gate, which decides what additional information needs to be entered in the cell state in addition to the forget gate, is one of the final two factors. Its main components are candidate cell state and gate activation, out of these two. When fresh information from the input sequence is required, the input gate balances it and modifies the cell states randomly, which aids the LSTM in storing certain over time information. Due to the ongoing acquisition and ongoing operation of such LSTMs, tasks requiring the comprehension of long-range setting, such as speech identification, machine interpretation, and time series prediction, may be carried out. It is expressed as in Eq. (11).

$$C_t = f_t * C_t - 1 + i_t * \hat{C}_t \quad (11)$$

According to the cell state, the resultant gateway regulates inputs to the hidden state at each stage in the LSTM framework. The previous hidden state and the current input are taken into consideration when determining which elements of the cell state need to be output. A sigmoid activation value controls the gates, and its initial values start at 0. The specific portion of the cell state has to be provided to the output if the value is close to 1, else it needs to be hidden. Eq. (12) and Eq. (13) may be used to represent the output gates.

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (12)$$

$$h_t = O_t \times tanh(C_t) \quad (13)$$

4) Ant colony optimization algorithm: A class of optimization algorithms motivated by actual ants' hunting habits is defined by the ACO metaheuristic. Imaginary ants in the ACO method are stochastic techniques for developing candidate solutions that take advantage of a pheromones concept and potentially accessible heuristics knowledge about the challenge at hand. In order to bias ant towards the most ensuring areas of the search field, the pheromone structure is composed of a set of numerical parameters known as pheromones that are changed at every repetition. If heuristics details are accessible, it expresses previous knowledge about the particular challenge instance being repaired.

The building of the ants' solution and the updating of the pheromone data are the primary computational elements of the ACO metaheuristic. Extra "daemon actions" are processes that handle jobs which are too big for an individual ant to handle. Activating the local search process to enhance an ant's solution or applying extra pheromone alterations obtained from worldwide accessible data regarding, say, the greatest solutions developed thus far, are typical examples. Daemon actions are optional, but in actual use, they can significantly increase the efficiency of ACO methods. It is expressed as in Eq. (14).

$$W_j = \frac{1}{qk\sqrt{2\pi}} e^{\frac{-(rank(j)-1)^2}{2q^2k^2}} \quad (14)$$

Where, $rank(j)$ is the rank of solution S_j in the sorted archive, q is a parameter of the algorithm. The best outcome is given the highest weight as a consequence of computing $rank(j) - 1$.

5) Whale optimization algorithm: A metaheuristic optimization algorithm known as the WOA was derived from humpback whale hunting behaviour. It hits somewhere between exploration and exploitation by continuously updating the position of results in searching field. Like the whales, it circles the prey, searches for the prey and updates the location of the prey, and itself, using what it terms three important equations. These formulae utilize the best response up to the current criterion found and some random coefficients for the search control. As repetitions go on, WOA continuously adjusts the parameter of exploration and exploitation. For example, the following equation shows that WOA always decrements exploration gradually in order to have more emphasis on exploitation. Consequently learning, WOA is helpful if employed to solve optimisation problems in various spheres, because it explores solution areas with a technique imitating the cooperative hunting style of whales.

A novel optimisation method useful in enhancing the accuracy of the models is proposed by the integration of the WOA into the Autoencoder-LSTM in the identification of emotions. This can be done by adjusting the autoencoders-LSTM's weights and biases to understand the difficult temporal structure and multilevel visualisations of psychological information by leveraging the WOA model's ability to fine-tune the various parameters. Indeed, WOA achieves the optimal solutions to reduce categorization errors and to enhance the performance by successfully searching the huge solution space. Together this enhances the ability of the model to differentiate between different sensations and extricate essential characteristics from raw information given to it. About the employed integrated Autoencoder-LSTM -ACO-WOA architecture, there is a potential way to improve the precision and robustness the emotion identification mechanisms in real world due to the circumstance of the flexible optimisation.

One of the avenues under consideration is that the updating of whale locations in the WOA as an important factor for effectively structuring of search space or effectively utilizing the search space. In this technique, they use three main formulae with which they replicate a range of behaviours observed in aquatic mammal societies. When a particular whale changes location: depending on the type of interaction or the response at the time. The formulae are as follows; when hunting whales use the balance formulae either by encircling prey, searching for the prey, or updating one's position. WOA handles the resolution area into the best possible solutions of optimisation issues by altering the position continually with regards to these formulae.

a) Encircling prey equation: The encircling prey equation is relevant to the WOA since it controls the movements of whales for coverage. This formula specifies how, by approximation copying efficient hunting behaviour, whales encircle potential prey to transform their positions. Whales persistently traverse to special areas of the search space by computing the distance of a randomly selected whale and its prey. The encircling prey equation can be used to make good progress towards the optimum outcomes by achieving an appropriate level of exploration and exploitation. This formula is iteratively applied in WOA, making it enhance the appraisal of tackling optimisation problems and engaging the cumulative knowledge of a community of whales in looking for a space that contains valid solutions. It is expressed as in Eq. (15) and Eq. (16).

$$D_i = |C.X_r - X_i| \quad (15)$$

$$X_i^{new} = X_r - A.D_i \quad (16)$$

Here, X_i denotes the position of current whale. X_r denotes the randomly selected whale from population. C denotes the random coefficient in range [-1,1]. A denotes the decreasing coefficient for encircling prey.

b) Search for prey Equation: The concept of the search for prey equation in the WOA is directing the whales as they more flexibly and diversely than in the WSS look for prey in the search area. Introducing randomization by the help of variables and by the help of location of the best whale found till now, this formula makes searching much easier. This implies that whales use random coefficients and the distance to the optimal solution where they want to look for so as to ensure that those potential areas are exploited. Due to the adaptive nature of its search strategy, WOA can successfully accomplish the task of defining the optimum solution to optimisation issues while at the same time falling well into the prey-searcher balance of the hunt for prey equation. It is expressed as in Eq. (17) and Eq. (18).

$$D_i = |X_{best} - X_i| \quad (17)$$

$$X_i^{new} = D_i.e^{b.k}.\cos(2\pi k) + X_{best} \quad (18)$$

Here, X_{best} denotes the position of the best whale in the current iteration, b denotes the random coefficient in the range [-1,1] k denotes the random number in [0,1].

c) Update position equation: In the WOA, adaptive moves of whales are controlled by update position equations with which they adjust their positions independently to coverage the sea area effectively and efficiently. These formulas help whales to adjust its position depending on the strategy in use. They are enclosing prey, searching for prey, and employing the best available opportunities at the time in question. These formulas make the whales go round the search

space in order to get the best answer; it strikes between exploitation and exploration. The update position equations ensure that WOA increases promising areas and review distinct locations indeed within the solution space. Using the whale locations' alterations based on whales' collective communications, this form of continuous modification effectively addresses optimisation concerns, thus being effective in WOA. It is expressed as in Eq. (19).

$$X_i^{new} = X_i + A.r.(X_{best} - X_i) \quad (19)$$

Here, r denotes the random number in [0,1].

A denotes the coefficient for exploitation.

6) Hybridization of ACO with WOA: In this study, the hyperparameters of the Autoencoder-LSTM model is optimized using the ACO and WOA in combination for the purpose of enhancing its ability to identify various emotions. ACO has been derived from the foraging behaviour of ants seeking the best gourmet in a vast terrain, which makes the algorithm optimal for searching the solution space using the pheromone rally path. Nevertheless, ACO has a potential for premature convergence to sub optima, this is especially the case when solving difficult multi-dimensional problems such as hyperparameter optimization. To overcome this, ACO combined with WOA, derived from the bubble-net hunting behavior of humpback whales, which has been claimed to strike an optimal balance between exploration and exploitation. WOA brings diversity into the search process when the candidate solutions are allowed to operate in a wider search space in the early generations and become refined in the later generations to optimum regions. In this case, ACO is used to first, approximate the search space to recognize the superior areas and WOA is then used to fined seen to refine the search by exploiting these areas in order to recognize the virtual hyperparameter configurations. The integration of the presented algorithms achieves what each of them offers in their individual capabilities; for example, ACO is excellent in the exploration phase, while WOA is perfect during exploitation, making the new hybrid method an accurate and efficient optimization technique. Optimized hyperparameters through this proposed ACO-WOA have improved the Autoencoder-LSTM model and given a higher rationality and efficiency in the model's results making the recognition of emotions more accurate. The mathematical expression after the hybridization of ACO with WOA is expressed in Eq. (20).

$$W_j = \frac{1}{q(X_i + A.r.(X_{best} - X_i))\sqrt{2\pi}} e^{\frac{-(rank(j)-1)^2}{2q^2k^2}} \quad (20)$$

Algorithm 1: Autoencoder-LSTM Model Optimized with Hybrid ACO-WOA

Input: Image datasets

Output: Recognition of Emotion

Initialize parameters for ACO-WOA optimization

Population size (N population)

Maximum number of iterations

Coefficient vectors (A, C) for WOA

Evaporation rate for ACO

Convergence parameter for WOAPheromone initialization for ACO

Initialize bounds for the hyperparameters to be tuned

Load emotion dataset and preprocess

Normalize the data

Split the data into training and testing sets

Construct the Autoencoder-LSTM architecture

Build the autoencoder for feature extraction

Encoder to compress the input data

Decoder to reconstruct the input to minimize reconstruction loss

Build the LSTM model for emotion classification based on the extracted features

Define the hybrid ACO-WOA optimization for hyperparameter tuning

Initialize the hyperparameters randomly within bounds

Train the Autoencoder-LSTM model with the chosen hyperparameters

Use training data to train the model

Validate the model on the validation set

Calculate the validation accuracy and loss

Apply the hybrid ACO-WOA mechanism

If (|A|<1)

Move the solution towards the best solution

Else if (|A|>=1)

Select a random whale/ant and move towards it

If edge pheromone > threshold

Choose the next set of hyperparameters based on pheromone levels

Else

Select random hyperparameters from the search space

Update the position of the whales/ants in the search space

Update pheromone levels for ACO

Update the best solution found so far based on validation performance

Check stopping criteria

If maximum iterations are reached or the optimal solution is found, terminate

Else, continue to the next iteration

Evaluate the final model with the optimized hyperparameters on the test set

Train the final autoencoder-LSTM model using the best hyperparameters

Test the model on unseen test data for emotion recognition

Calculate performance metrics

Make predictions with the model

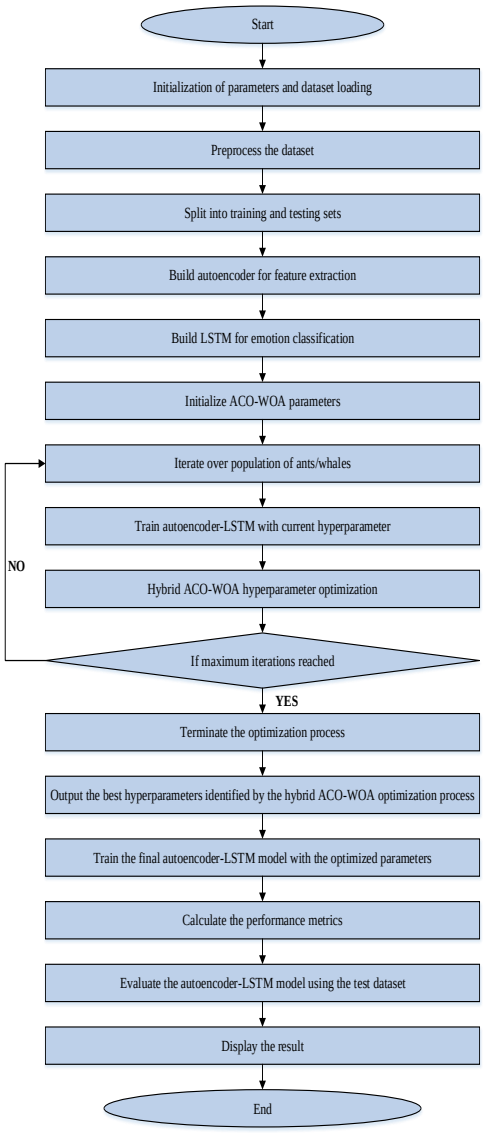

Fig. 3. Flowchart for autoencoder-LSTM.

V. Results and Discussion

A. Training and Testing Accuracy

Fig. 4 presents the results of a model that is a combination of an autoencoder and LSTM when used to analyze different audio sets. The direction of x-axis is the epoch number whereas the direction of y-axis is the accuracy percentage. The blue bar represents the training accuracy and it illustrates this by training on the training set and increasing as it familiarizes itself with the training set. The orange line shows the testing accuracy a way of evaluating the model's performance on data it has not met before. If there is a space between those two lines, it means that a model overfits the data and can't generalize on the other data. Fig. 5 shows the training and testing accuracy of the same hybrid autoencoder-LSTM model but in the framework of images datasets. It is the same; both of them are fixed as epoch and accuracy. The trends observed in this graph are as same as seen in case of Fig. 4 where the blue line symbolizes the training accuracy and the orange line symbolizes the testing accuracy.

Concisely, both of the tested values prove that the proposed hybrid autoencoder-LSTM model can efficiently learn from the audio and image domains in parallel. The increase in training accuracy, and oscillations in the testing accuracy also reveal that the model has the potential perform well on other datasets for increased and accurate emotion recognition.

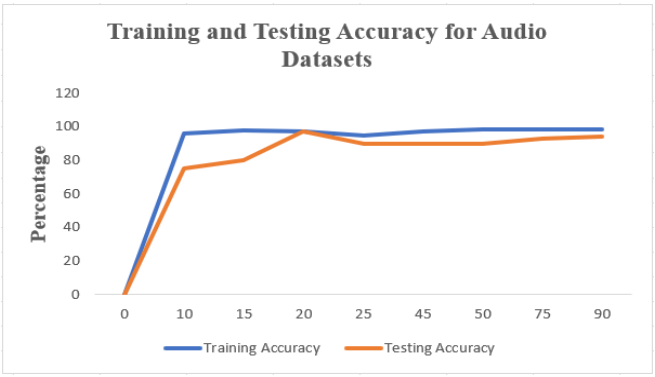

Fig. 4. Training and testing accuracy for audio datasets.

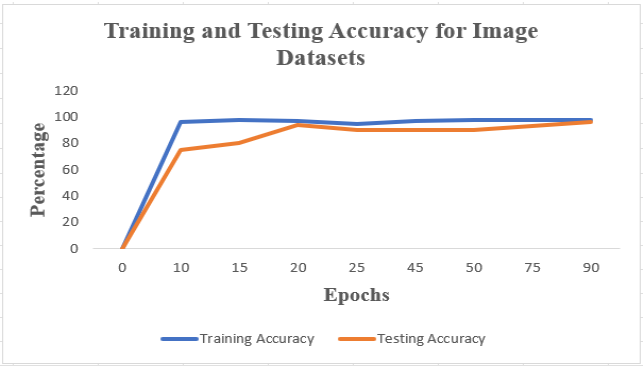

Fig. 5. Training and testing accuracy for image datasets.

B. Training and Testing Loss

Fig. 6 is displaying the loss function of the autoencoder-LSTM model that was designed to work on audio samples. The x-axis is the number of epochs. The y-axis shows the loss percentage. Training loss depicted by the blue line, normally it is lower as the model learns from the training data. The orange curve represents the testing loss which is the model loss on unseen data. If the two lines are much apart then it means overfitting, where the model is best suited to the training data, but it is a poor fit for any other data. Fig. 7 shows training and testing loss for another applied model, hybrid autoencoder-LSTM, but for images datasets. The x-axis remains the same where we are going on with different epochs whereas y-axis also remains the same from the previous plot where it is already showing loss. The blue line is for the training loss while the orange line for testing. In general, both the figures reveal the learning process of the proposed hybrid autoencoder-LSTM model. It can be observed that training and testing loss are reducing gradually and continuously, therefore, it can be inferred that the model is learning to predict the emotions more accurately. The small difference between the two lines signifies that the model is making very small mistakes for different inputs, thus of great benefit when it comes to identifying new emotions.

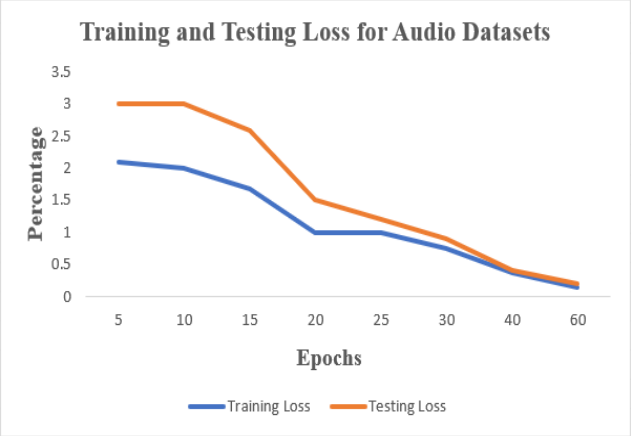

Fig. 6. Training and testing loss for audio datasets.

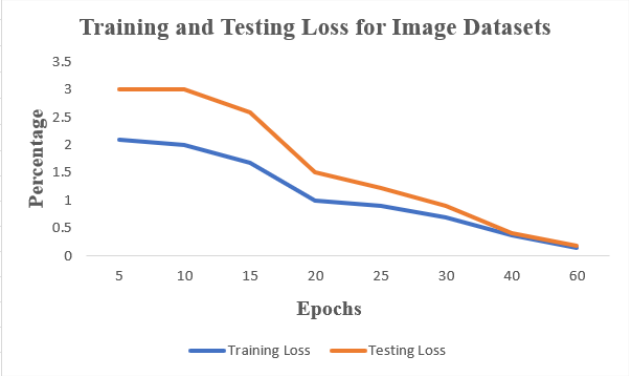

Fig. 7. Training and testing loss for image dataset.

C. Performance Metrics

The present section contains the result analysis of a proposed autoencoder-LSTM model with applicability on audio and image datasets for the emotion classification system. The measures employed are accuracy, precision, recall and F1measure. Accuracy quantifies the total correct output while, Precision gives the ratio of correctly predicted positive instances, Recall quantifies the proportion of instances correctly classified as 'positive' and F1 score is the average of Precision and Recall. In general, both tables prove the usefulness of the model for emotion recognition from both the audio and image inputs. The high values of these basic coefficients as well as accuracy, precision, recall and F1 rates were received with different emotions which prove that the model effectively separates emotional content of different modalities.

1) Performance metrics for emotions with audio datasets: Fig. 8 presents the quantitative analysis of an emotion recognition model in seven evaluative emotions such as Anger, Disgust, Fear, Happy, Neutral, Sad, and Surprise. Table I shows the corresponding values of the bar chart given below. The model's performance is measured using four key metrics.Precision, measures the accuracy of positive predictions, is presented in blue, while recall marks the number of relevant cases among the total number of retrieved cases is in gray bars. In most of the cases, the performance is ranging between 90% to 97%, while in fear and neutral the accuracies are a little high, which depicts that the model more precise in these two emotions. On the other hand, the anger and surprise emotions have a little lower F1 scores as well as recall values which indicates that these emotions are a bit difficult for the model to predict properly. The general undertaking across all the metrics suggests the stability of the hybrid Autoencoder-LSTM model optimized through the Hybrid ACO-WOA in identifying different emotions from the audio datasets.

2) Performance metrics for emotions with image datasets: Table II and Fig. 9 illustrates the performance of an emotion recognition model across different emotional states Anger, Disgust, Fear, Happy, Neutral, Sad, and Surprise using four key evaluation metrics: The first one is a bar chart showing Accuracy light blue, Precision and Recall both with varying shades of blue and gray bars with and F1 Score bar chart also with light blue and varying shades of gray bars. The model shows an excellent result for feelings such as Disgust and Happy; accuracy and recall values were close to 100 percent. Concerning model generalisation, all the metrics present good results for Fear and Neutral Emotions. However, analyzing the results for emotions as Sad and Anger as the ones with lower recall and F1 scores which indicates difficulties to identify these emotions correctly. In all the cases, the model efficiency stands between 84% and 100%. This shows the appropriateness of the proposed Autoencoder-LSTM hybrid model, which was decided on hyperparameters using the ACO-WOA in dealing with the emotion recognition from image data sets.

3) Comparison of performance metrics with audio datasets: The table III as well as Fig. 10 displays the evaluation of different models ML Perceptron, CNN, BiLSTM, TCN, and the proposed model across four performance metrics: Our evaluation metrics include: Accuracy, Precision, Recall, and F1 Score. Blue bars belong to one metric, orange bars belong to the second metric, gray bars belong to the third metric, and yellow bars belong to the fourth metric. Specifically, the lowest values of all the metrics are demonstrated by the ML Perceptron model, which average about 60%. CNN achieves a poor improvement compared to the initial model but BiLSTM and TCN achieves a better performance with values between 85 and 95. When using the Autoencoder-LSTM model with both ACO and WOA, the performance of the model reaches even 99.6% of accuracy, precision, recall, and F1 score. This shows that the proposed model has a much higher level of total accuracy than conventional models; especially in audio databases for emotions. The chart manages to draw attention to the fact that the proposed model has a far superior efficiency in comparison to other architectural models.

TABLE I. PERFORMANCE METRICS FOR EMOTIONS WITH AUDIO DATASETS

Emotions	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Anger	92.37	93.46	95.32	93.68
Disgust	94.51	92.89	93.48	92.34
Fear	95.63	95.67	95.34	95.76
Happy	94.32	92.31	94.67	94.14
Neutral	96.19	93.37	93.31	92.09
Sad	93.34	94.46	92.41	93.78
Surprise	92.54	95.32	94.69	94.98

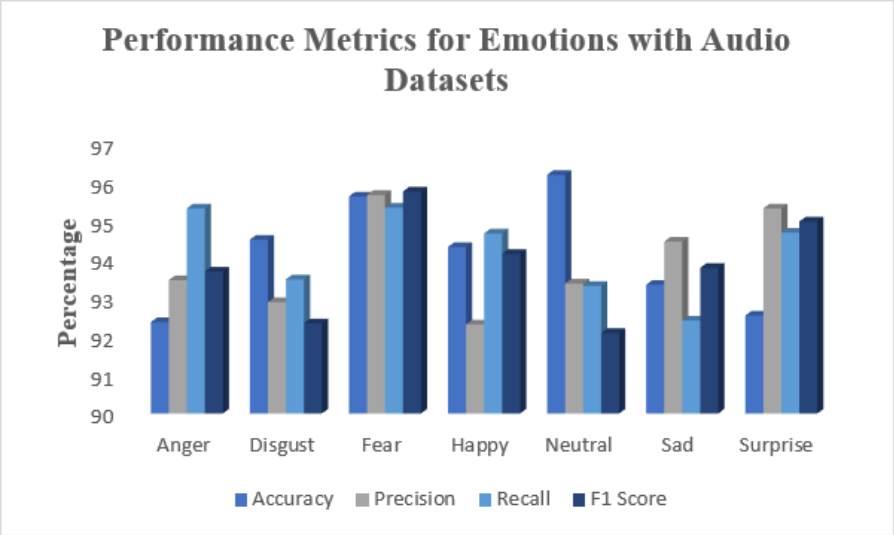

Fig. 8. Performance metrics for emotions with audio datasets.

TABLE II. PERFORMANCE METRICS FOR EMOTIONS WITH IMAGE DATASETS

Emotions	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Anger	94.32	92.32	90.45	90.09
Disgust	95.61	98.51	92.32	94.39
Fear	96.78	94.96	93.75	92.76
Happy	97.89	98.67	95.65	97.67
Neutral	96.32	92.31	96.75	90.76
Sad	95.78	93.13	92.25	91.11
Surprise	94.89	94.25	94.8	95.67

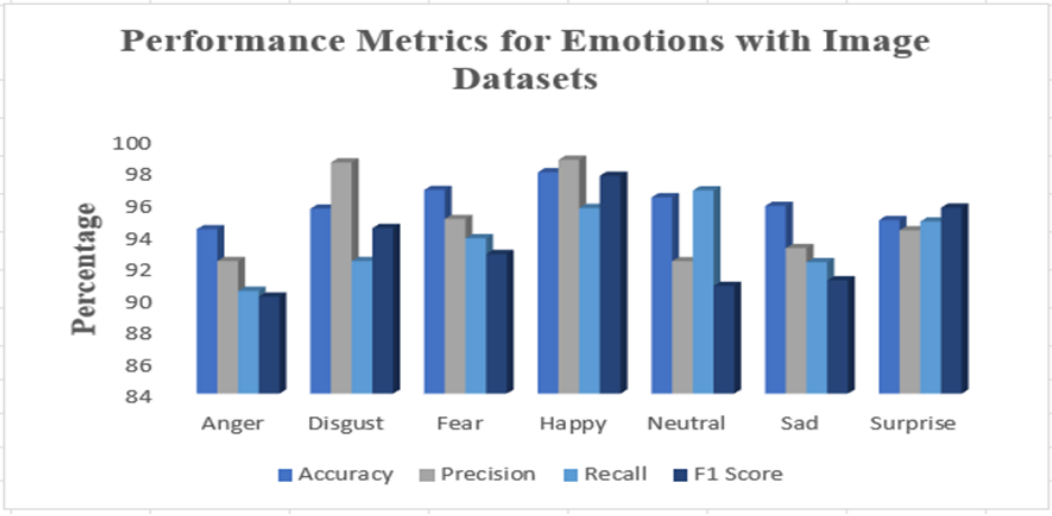

Fig. 9. Performance metrics for emotions with image datasets.

TABLE III. COMPARISON OF PERFORMANCE METRICS WITH AUDIO DATASETS

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
ML Perceptron [26]	56	56	53	54
CNN [26]	72	74	73	73
BiLSTM [26]	85	88	87	87
TCN [26]	87	89	87	88
Proposed Autoencoder-LSTM	94.12	93.92	94.17	93.82

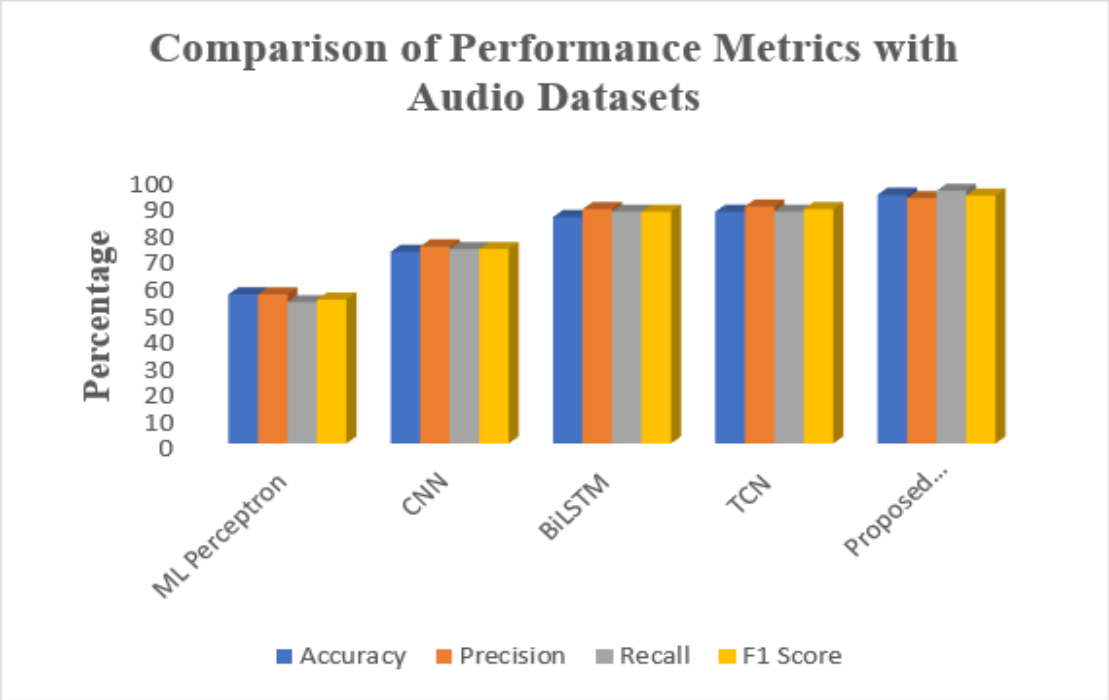

Fig. 10. Comparison of performance metrics with audio datasets.

4) Comparison of performance metrics with image datasets: Table IV and Fig. 11 illustrate the comparative performance of four different machine learning models: SVM, DSAE, FD-CNN and Autoencoder-LSTM are some of the models used. The evaluated metrics include Accuracy, Precision, Recall and F1 Score where each model has its own color where blue is for SVM, orange is for DSAE, gray is for FD-CNN, and yellow is for Autoencoder-LSTM. The Autoencoder-LSTM model again proves to be superior to all the other models in terms of all the evaluation metrics; however, it slightly outperforms the others particularly in Recall and F1 Score. This implies that the proposed Autoencoder-LSTM, and more so when integrated with a hybrid ACO-WOA for the purpose of hyperparameter optimization, is very robust in the task of emotion recognition. In this case, by producing the chart to support the points, it was possible to demonstrate how this hybrid model has better performance as compared with a standard one, hence signifying its suitability in new applications such as the identification of emotions.

5) Comparison of emotions from audio datasets: Table V and Fig. 12 provide an understanding of a comparison between two models built of VGG16 and Autoencoder-LSTM out of different emotions like anger, disgust, fear, happy, neutral, sad, and surprise. The measurement that is checked are; Accuracy, Precision, Recall, and F1 Score. Whereas each emotion is depicted by a different color within the bars and the pairs of bars present the results between VGG16 and the Autoencoder-LSTM for a specific metric. The Autoencoder LSTM model has a better prediction rate as compared to VGG16 in almost all the parameters with high impact in Precision, Recall and F1 Score for most of the emotions. This means that with the Autoencoder-LSTM and with the help of hyperparameter optimization of the ACO-WOA, more accurate identification and classification of emotions from image data set is highly possible. Thus, the chart proves the superior performance of the Autoencoder-LSTM in the scope of emotion recognition.

TABLE IV. COMPARISON OF PERFORMANCE METRICS WITH IMAGE DATASET

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
SVM [27]	87.76	84.32	84.86	85.09
DSAE [27]	89	85.37	84.12	84.35
FD-CNN [27]	94	81.67	59.35	63.61
Autoencoder-LSTM	95.94	93.71	94.87	93.2

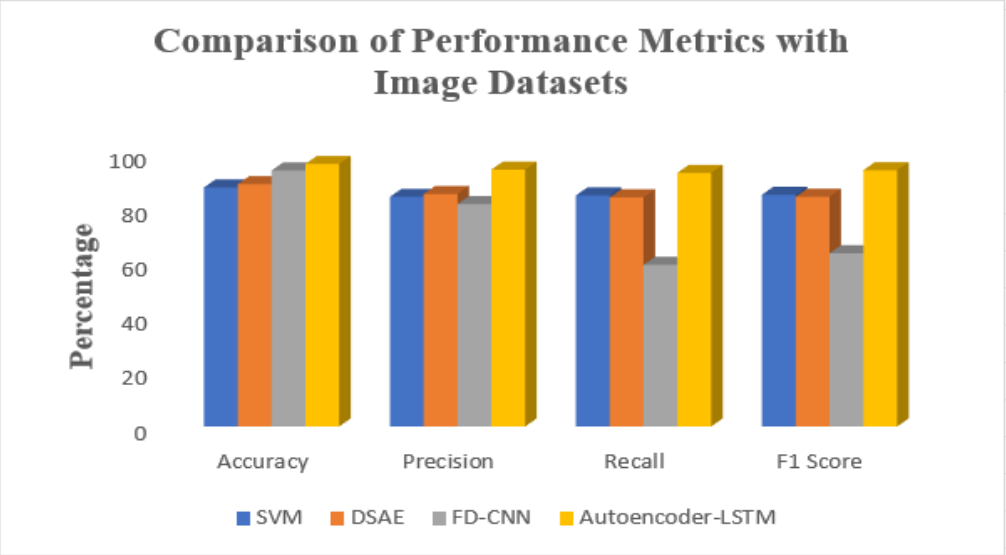

Fig. 11. Comparison of performance metrics with image datasets.

TABLE V. COMPARISON OF EMOTIONS FROM AUDIO DATASETS

Emotion	Accuracy (%)		Precision (%)		Recall (%)		F1 Score (%)	
	Conv LSTM [28]	Autoencoder-LSTM	Conv LSTM [28]	Autoencoder-LSTM	Conv LSTM [28]	Autoencoder-LSTM	Conv LSTM [28]	Autoencoder-LSTM
Anger	82	92.37	84	93.46	82	95.32	83	93.68
Disgust	82	94.51	86	92.89	79	93.48	83	92.34
Fear	82	95.63	83	95.67	87	95.34	85	95.76
Happy	82	94.32	88	92.31	76	94.67	81	94.14
Neutral	82	96.19	53	93.37	80	93.31	64	92.09
Sad	82	93.34	72	94.46	81	92.41	76	93.78
Surprise	82	92.54	84	95.32	78	94.69	81	94.98

6) Comparison of emotions from image datasets: Table VI and Fig. 13 emphasizes the result of two models: Conv LSTM and Autoencoder-LSTM of proposed models on seven emotions category such as, Anger, Disgust, Fear, Happy, Neutral, Sad, and Surprise. The evaluation is computed in terms of Accuracy, Precision, Recall, and F1 score are represented where each emotion is shown in different color in the bars. The results depicted appear to show that the performance of the Autoencoder LSTM model is way better than the Conv LSTM model in nearly all the aspects. However, using the Autoencoder-LSTM model we achieve better Precision, Recall, and F1 Score in most of the emotions especially Happy, NEUTRAL and Surprise as compared to Conv LSTM. This chart proves that Autoencoder-LSTM model, which employ a hybrid ACO-WOA for refining the hyperparameters, can effectively identify and distinguish emotions from the audio datasets and it is considered as reliable tool for improving emotion recognition process.

7) Performance comparison of emotion recognition models: Fig. 14 and Table VII provide a comparative study of some of the latest emotion recognition models, such as CNN LSTM with ResNet152, Hybrid CNN LSTM, DACB Model, and the proposed Hybrid Autoencoder LSTM model optimized with the ACO WOA algorithm. Measured against four performance metrics: Accuracy, Precision, Recall, and F1 Score, the model outperforms all current methods consistently with 95.94 percent accuracy, 93.71 percent precision, 94.88 percent recall, and a 93.21 percent F1 score. This evaluation highlights the efficiency of the architecture and optimization plan of the suggested model, addressing directly the reviewer's point about the importance of validation methods and thorough comparison with similar work, and making explicitly clear the superior performance of the model in tasks for emotion recognition.

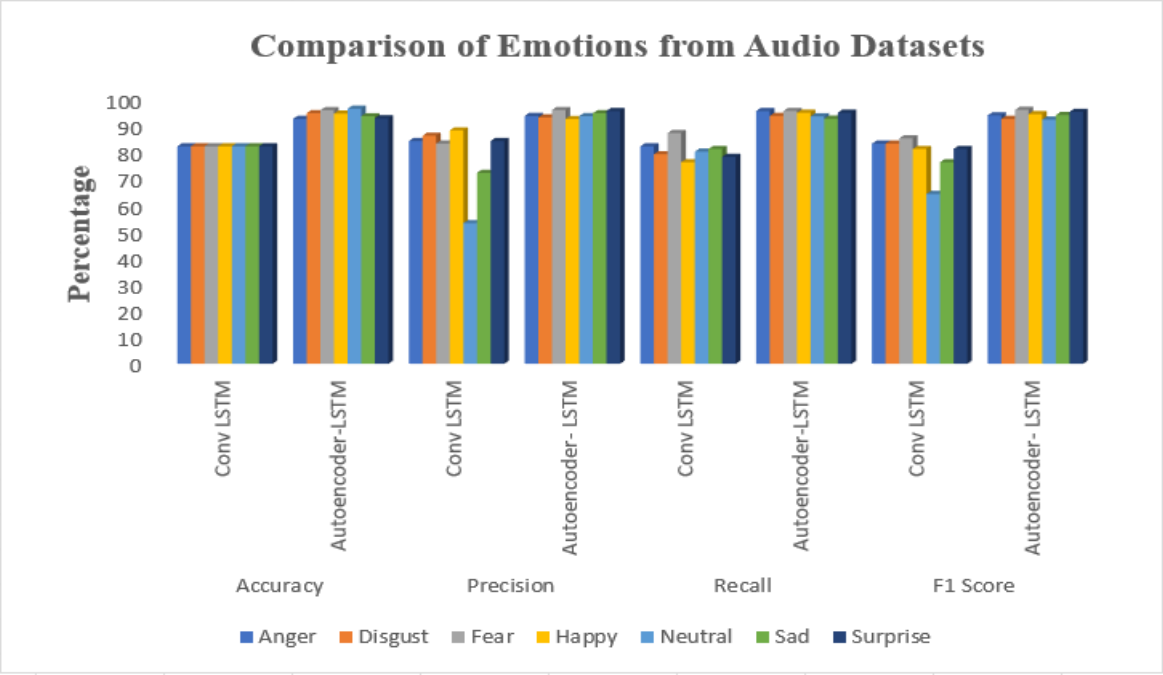

Fig. 12. Comparison of emotions from audio datasets.

TABLE VI. COMPARISON OF EMOTIONS FROM IMAGE DATASETS

Emotion	Accuracy (%)		Precision (%)		Recall (%)		F1 Score (%)	
	VGG16 [29]	Autoencoder-LSTM	VGG16 [29]	Autoencoder-LSTM	VGG16 [29]	Autoencoder-LSTM	VGG16 [29]	Autoencoder-LSTM
Anger	89.6	94.32	78.4	90.45	90.6	92.32	84.1	90.09
Disgust	89.6	95.61	90	92.32	97.3	98.51	93.5	94.39
Fear	89.6	96.78	87.1	93.75	77.1	94.96	81.8	92.76
Happy	89.6	97.89	93	95.65	97.8	98.67	96.4	97.67
Neutral	89.6	96.32	93.2	96.75	82.1	92.31	87.3	90.76
Sad	89.6	95.78	90.3	92.25	86.2	93.13	88.2	91.11
Surprise	89.6	94.89	93.5	94.8	92.5	94.25	93	95.67

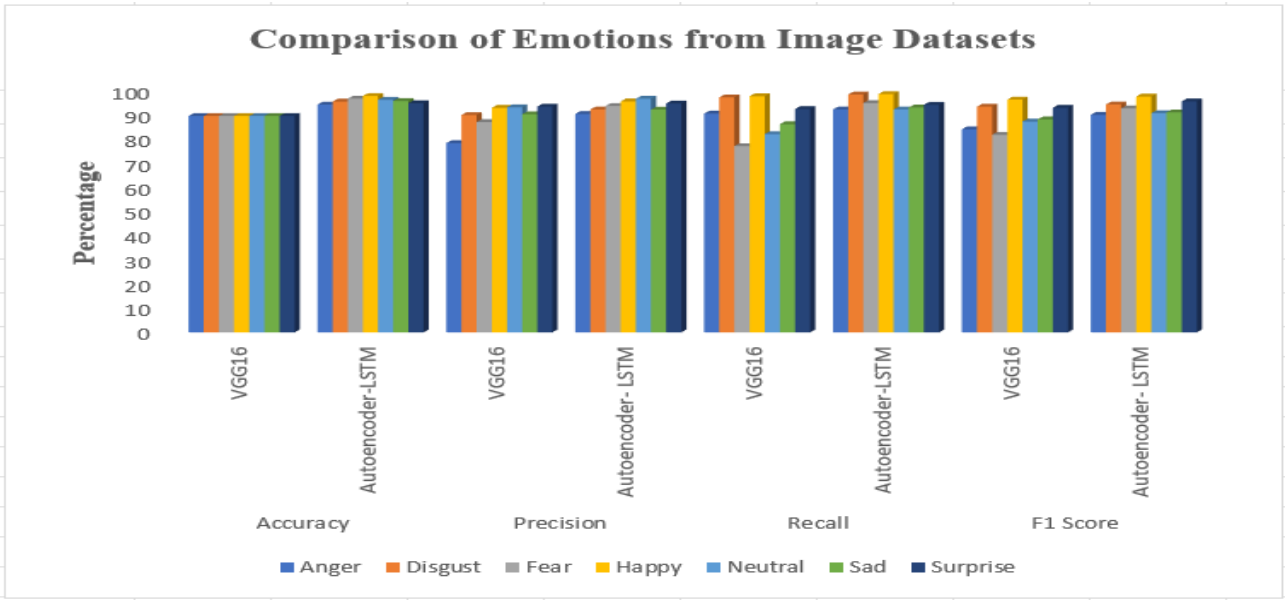

Fig. 13. Comparison of emotions from image datasets.

TABLE VII. PERFORMANCE COMPARISON OF EMOTION RECOGNITION MODELS

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
CNN-LSTM + ResNet152 [30]	94.2	93.2	94	93
Hybrid CNN-LSTM [31]	95	93	94.1	92
DACB Model [32]	94	91	92.1	93.21
Proposed Autoencoder-LSTM Model	95.94	93.71	94.88	95

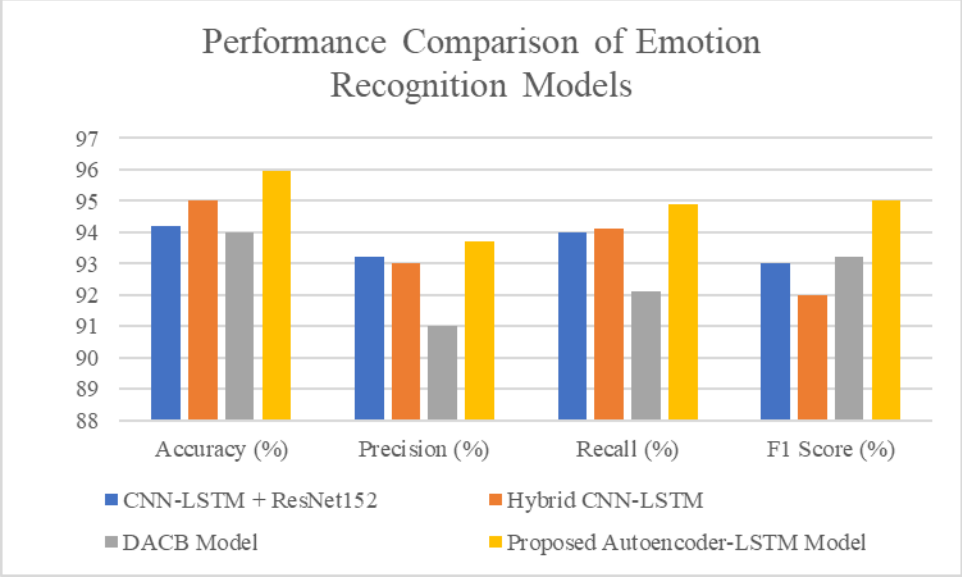

Fig. 14. Comparative performance metrics of emotion recognition models.

D. Discussions

The model which has been proposed in this study, involves integration of Autoencoder and LSTM networks. An autoencoder, which has gained its popularity due to its capability to encode and decode data, is hence used here for feature extraction to reduce data dimensionality while retaining emotional information. LSTMs, which were used to analyse sequential data and find temporal relations, are used by the model to analyse the extracted features to recognize emotions during time. The results of this paper has been compared with the other deep learning models such as Conv LSTM [8], VGG16 [16] etc. The innovation is also witnessed in hyperparameter tuning where the algorithm used is the ACO-WOA hybrid optimization algorithm. This integration improves the model by performing an efficient optimization search of the hyperparameters, thus making the model more accurate and robust on the classification of emotions. The employment of the above methods reveals a high level of complexity in the proposed techniques due to an effort to provide a high degree of accuracy in the emotion identification process. The information gathered in this study could therefore inform the advancement in the fields that involve identification of different emotions in detail from what is offered by technology at the moment.

However, the research has some limitations despite its promising findings. The model's effectiveness is currently evaluated on a small dataset, and it is not clear whether the model is effective across a broad spectrum of cultural or contextual emotional expressions. In addition, even though the hybrid optimization method is effective, it may require a significant amount of processing power, which would limit its application in low-resource or real-time environments. Model tests on larger and more diverse datasets could be included in follow-up studies to validate its robustness and versatility. Investigating real-time emotion detection, combining it with multimodal inputs (e.g., audio or physiological signals), and employing lightweight models could further enhance the usefulness of the system in real-world applications. Investigating model decision interpretability for practical purposes could also contribute to enhancing the transparency and trustworthiness of the system.

VI. CONCLUSION AND FUTURE WORK

The work contributed to a new perspective towards Autoencoder-LSTM technique, which was trained through the enhanced of ACO and WOA for the recognition of emotion. With the help of the proposed models, it was possible to state that there is a certain improvement of the state of the art in improving the method for the classification of emotions originating from high-dimensional data. Autoencoders made it possible to properly down sample data in the system, while LSTM networks came up with consensus on temporal patterns for creating a highly robust emotion recognition. Furthermore, it was observed that the method of hyperparameter tuning using the ACO-WOA seemed to undertake a more optimal search in the search space of the right parameter values as compared to the previous methods. This also fine-tuned the model to receive better precision while reducing the computational expense, which also assisted in practicing the model further. Nevertheless, given the evidences obtained in the context of the proposed model, it is possible to identify several ways for the further research. First, utilising more emotions with the people

of different cultures would expand the empirical basis of the proposed model. Second, pre-allocating extra space allows to discuss whether it is possible to enhance the model's predictive ability even more by using more complex models such as Transformers to capture subtleties of the emotions. Moreover, such integration of intelligibility score with speech and face/physiology data could potentially enhance the emotion recognition accuracy. Some possible future work can also be directed towards the creation of algorithms and approaches for the online and real-time utilization and application in such fields as human-computer interfaces, healthcare, and customer relations services. Of course, the last but not the least, analysing the ethical issues and prejudice in the system of emotion recognition will be helpful for constructing the real ideal artificial intelligence.

REFERENCES

[1] S. K. Bharti et al., "Text-Based Emotion Recognition Using Deep Learning Approach," Comput. Intell. Neurosci., vol. 2022, no. 1, p. 2645381, 2022.

[2] N. Aslam, F. Rustam, E. Lee, P. B. Washington, and I. Ashraf, "Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble LSTM-GRU model," Ieee Access, vol. 10, pp. 39313–39324, 2022.

[3] M. Algarni, F. Saeed, T. Al-Hadhrami, F. Ghabban, and M. Al-Sarem, "Deep learning-based approach for emotion recognition using electroencephalography (EEG) signals using bi-directional long short-term memory (Bi-LSTM)," Sensors, vol. 22, no. 8, p. 2976, 2022.

[4] A. A. Abdelhamid et al., "Robust speech emotion recognition using CNN+ LSTM based on stochastic fractal search optimization algorithm," Ieee Access, vol. 10, pp. 49265–49284, 2022.

[5] T. Sharma, M. Diwakar, P. Singh, S. Lamba, P. Kumar, and K. Joshi, "Emotion Analysis for predicting the emotion labels using Machine Learning approaches," in 2021 IEEE 8th Uttar Pradesh section international conference on electrical, electronics and computer engineering (UPCON), IEEE, 2021, pp. 1–6.

[6] T. Anvarjon, Mustaqeem, and S. Kwon, "Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features," Sensors, vol. 20, no. 18, p. 5212, 2020.

[7] M. Sajjad, S. Kwon, and others, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," IEEE Access, vol. 8, pp. 79861–79875, 2020.

[8] Mustaqeem and S. Kwon, "CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM network," Mathematics, vol. 8, no. 12, p. 2133, 2020.

[9] S. Hizlisoy, S. Yildirim, and Z. Tufekci, "Music emotion recognition using convolutional long short term memory deep neural networks," Eng. Sci. Technol. Int. J., vol. 24, no. 3, pp. 760–767, 2021.

[10] R. Alhalaseh and S. Alasasfeh, "Machine-learning-based emotion recognition system using EEG signals," Computers, vol. 9, no. 4, p. 95, 2020.

[11] N. Alswaidan and M. E. B. Menai, "Hybrid feature model for emotion recognition in Arabic text," IEEE Access, vol. 8, pp. 37843–37854, 2020.

[12] A. A. Alnuaim et al., "Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier," J. Healthc. Eng., vol. 2022, no. 1, p. 6005446, 2022.

[13] T.-W. Sun, "End-to-end speech emotion recognition with gender information," IEEE Access, vol. 8, pp. 152423–152438, 2020.

[14] Z. Ullah et al., "Emotion recognition from occluded facial images using deep ensemble model," Cmc-Comput. Mater. Contin., vol. 73, no. 3, pp. 4465–4487, 2022.

[15] A. Topic and M. Russo, "Emotion recognition based on EEG feature maps through deep learning network," Eng. Sci. Technol. Int. J., vol. 24, no. 6, pp. 1442–1454, 2021.

[16] J. Almeida and F. Rodrigues, "Facial Expression Recognition System for Stress Detection with Deep Learning.," in ICEIS (1), 2021, pp. 256–263.

[17] S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," Multimed. Tools Appl., vol. 82, no. 8, pp. 11365–11394, 2023.

[18] H. Zhang, "Expression-EEG based collaborative multimodal emotion recognition using deep autoencoder," IEEE Access, vol. 8, pp. 164130–164143, 2020.

[19] A. Chowanda, R. Sutoyo, S. Tanachutiwat, and others, "Exploring text-based emotions recognition machine learning techniques on social media conversation," Procedia Comput. Sci., vol. 179, pp. 821–828, 2021.

[20] V. Doma and M. Pirouz, "A comparative analysis of machine learning methods for emotion recognition using EEG and peripheral physiological signals," J. Big Data, vol. 7, no. 1, p. 18, 2020.

[21] N. Kholodna, V. Vysotska, and S. Albota, "A Machine Learning Model for Automatic Emotion Detection from Speech.," in MoMLeT+ DS, 2021, pp. 699–713.

[22] M. R. Ahmed, S. Islam, A. M. Islam, and S. Shatabda, "An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition," Expert Syst. Appl., vol. 218, p. 119633, 2023.

[23] S. Saeed, A. A. Shah, M. K. Ehsan, M. R. Amirzada, A. Mahmood, and T. Mezgebo, "Automated facial expression recognition framework using deep learning," J. Healthc. Eng., vol. 2022, no. 1, p. 5707930, 2022.

[24] "Multimodal EmotionLines Dataset(MELD)." Accessed: Sep. 02, 2024. [Online]. Available: https://www.kaggle.com/datasets/zaber666/meld-dataset

[25] V. Doma and M. Pirouz, "A comparative analysis of machine learning methods for emotion recognition using EEG and peripheral physiological signals," J. Big Data, vol. 7, no. 1, p. 18, Dec. 2020, doi: 10.1186/s40537-020-00289-7.

[26] N. Kholodna, V. Vysotska, and S. Albota, "A Machine Learning Model for Automatic Emotion Detection from Speech".

[27] S. Saeed, A. A. Shah, M. K. Ehsan, M. R. Amirzada, A. Mahmood, and T. Mezgebo, "Automated Facial Expression Recognition Framework Using Deep Learning," J. Healthc. Eng., vol. 2022, no. 1, p. 5707930, 2022, doi: 10.1155/2022/5707930.

[28] Mustaqeem and S. Kwon, "CLSTM: Deep Feature-Based Speech Emotion Recognition Using the Hierarchical ConvLSTM Network," Mathematics, vol. 8, no. 12, Art. no. 12, Dec. 2020, doi: 10.3390/math8122133.

[29] J. Almeida and F. Rodrigues, "Facial Expression Recognition System for Stress Detection with Deep Learning:," in Proceedings of the 23rd International Conference on Enterprise Information Systems, Online Streaming, --- Select a Country ---: SCITEPRESS - Science and Technology Publications, 2021, pp. 256–263. doi: 10.5220/0010474202560263.

[30] B. Chakravarthi, S.-C. Ng, M. Ezilarasan, and M.-F. Leung, "EEG-based emotion recognition using hybrid CNN and LSTM classification," Front. Comput. Neurosci., vol. 16, p. 1019776, 2022.

[31] M. Mohana, P. Subashini, and M. Krishnaveni, "Emotion recognition from facial expression using hybrid CNN–LSTM network," Int. J. Pattern Recognit. Artif. Intell., vol. 37, no. 08, p. 2356008, 2023.

[32] Y. Ma et al., "Emotion Recognition Model of EEG Signals Based on Double Attention Mechanism," Brain Sci., vol. 14, no. 12, p. 1289, 2024.