Library Progress International *Print version ISSN 0970 1052*
Vol.44 No.3, Jul-Dec 2024: P.12685-12696 *Online version ISSN 2320 317X*

Original Article *Available online at www.bpasjournals.com*

Multimodal Representation of Depression and Anxiety Severity Prediction Using Adaptive Capsule Network with Electric Eel Foraging Optimization

Dr. J. Venkatesh[1], K. Shantha Kumari[2], V. Rekha[3], Nalajam Geethanjali[4], S. K. Rajesh Kanna[5] and Dr. K. Sivakumar[6]

[1]Professor, Department of Computer Science and Engineering, Chennai Institute of Technology, Kundrathur, Chennai,India.
Orcid id: (0000-0002-4259-130X)
[2]Associate Professor, Department of Data Science and Business Systems, SRM Institute of Science and Technology, Kattankulathur, India
Orcid id: (0000-0001-5113-2183)
[3]Assistant Professor, Department of AI &DS, Panimalar Engineering College, Chennai, India
Orcid id: (0009-0008-7974-9853)
[4]Assistant Professor, Department of Artificial Intelligence and Machine Learning, Madanapalle Institute of Technology & Science, Kadiri Road, Madanapalle, Andhra Pradesh 517325
Orcid id: (0009-0007-3643-9540)
[5]Professor, Department of Mechanical Engineering, Rajalakshmi Institute of Technology, Chennai, India
Orcid id: (0000-0003-1013-008X)
[6]Professor, Department of Mechanical Engineering, P.T.LEE Chengalvaraya Naicker College of Engineering and Technology, Kancheepuram, India
Orcid id: (0000-0002-2647-5800).
[1]venkateshj@citchennai.net, [2]shanthak@srmist.edu.in, [3]rekhav20@gmail.com, [4]ngeethanjali.mits@gmail.com, [5]skrkanna@gmail.com and [6]shivakees@gmail.com

How to cite this article: J. Venkatesh, K. Shantha Kumari, V. Rekha, Nalajam Geethanjali, S. K. Rajesh Kanna K. Sivakumar (2024). Multimodal Representation of Depression and Anxiety Severity Prediction Using Adaptive Capsule Network with Electric Eel Foraging Optimization. *Library Progress International,* 44(3), 12685-12696.

ABSTRACT
The assessment of a person's mental state using data-driven models to gauge the severity of their symptoms is known as depression and anxiety severity prediction. This process aids in diagnosis, treatment planning, and therapy progress tracking. Predicting depression and anxiety severity from audio and video data is essential, and various techniques have been implemented to achieve this. However, the existing methods have lack of accuracy, precision and high error rate. To overcome the aforementioned problem, Adaptive Capsule Network (ACN) with Electric eel Foraging Optimization (ACN-EeFO) is proposed for accurately predicting depression and anxiety severity from audio, video data. In this input image is taken from two datasets such as AVEC2013 and AVEC2014 datasets. Then the video and audio data are extracted using Multi-Axis Vision Transformer (MaxViT) and Block-Based Haar Wavelet Transform (B-BHWT). Following that, the extracted audio and video data are fused using Efficient Long-range Attention Network (EL-AN). Then the classification is done by using Adaptive Capsule Network and optimized with Electric eel Foraging Optimization (ACN-EeFO) for classify the different stages of depression levels using BDI-II score. The introduced system is executed in python. The efficiency of the proposed ACN-EeFO is analyzed using 2 datasets and attains 99.82% accuracy, 99.23% F1-Score and attains better results compared with the existing methods. In the future, instead of merging the audio and video features at the end, it will attempt to find out if doing so enhances prediction performance. Additionally, scientists plan to utilize this framework for additional tasks including the prediction of various diseases.

Keywords: Prediction of depression and anxiety severity, Multi-Axis Vision Transformer, Block-Based Haar Wavelet Transform, Adaptive Capsule Network, Electric eel Foraging Optimization.

1. INTRODUCTION:

Anxiety depression is the most prevalent subtype of major depressive disorder (MDD), which affects over 350 million individuals globally. High levels of agitation and restlessness are characteristics of anxious depression, which also has worse treatment outcomes and a greater chance of recurrence. To improve results, identifying this subset is essential. Treatment personalization and targeting are thought to enhance results [1-3]. According to the DSM-5, depression is a type of mood disorder marked by severe depressive episodes that last for at least two weeks. These episodes are characterized by feelings of melancholy, diminished enjoyment, low self-esteem, irregular sleep and eating patterns, difficulty concentrating, and exhaustion. It affects 280 million people worldwide, or around 5% of adults [4-6]. Self-report or professional evaluation is frequently used to identify mental health disorders such as depression, anxiety, and suicide risk. Self-report measures require more time to administer under various settings, but they have good sensitivity. According to a 2017 meta-analysis, machine learning is advised for enhanced predictive capacities, arguing that present techniques are no more effective than chance [7-10]. The prevalence of mood disorders is rising quickly, which has an impact on people's quality of life. Mood disorders include major depressive disorders, bipolar disorder, dysthymia, and cyclothymia. According to the WHO, 87% of deaths worldwide are due to suicide accidents, and 3.8% of people worldwide suffer from depression. An estimated 40 million people have bipolar disorder [11-14]. Pronunciation, intonation, auditory content, and facial expressions are all linked to depression risk. Risk subjects minimize eye contact, make fewer facial gestures, and move their mouths less. Semantics and grammar in audio content are useful for detection. Deep learning-based neural network models offer benefits for both label prediction and feature extraction [15-16]. Patients' well-being can be continuously assessed thanks to the collection of physiological and behavioral data in real-time by mobile devices and sophisticated sensors. This reduces participant burden and recall bias while enabling better health outcomes, treatment adherence, and timely care. The creation of realistic models for the collection, processing, and analysis of behavioral sensor data is the subject of recent research, which will make it possible to build health monitoring systems for a variety of uses, such as tracking human activity, wellbeing, and early mental illness preventive strategies [17-18].

Novelty and Contribution

The Novelty and contribution of this paper is given below:

- In this manuscript, an Adaptive Capsule Network with Electric eel Foraging Optimization (ACN-EeFO) is proposed
- The features such as video and audio, are extracted using Multi-Axis Vision Transformer (MaxViT) and Block-Based Haar Wavelet Transform (B-BHWT)
- Feature fusion of extracted Video and audio are done using Efficient Long-range Attention Network (EL-AN)
- Classification are done using Adaptive Capsule Network (ACN) it is used for classify the depression and anxiety severity .And Optimization are done using Electric eel Foraging Optimization (EeFO) is used to optimize the weight parameters of ACN- EeFO for improving accuracy .

2. LITERATURE SURVEY:

In 2023, Sükei, et al [17] has introduced a Long Short-Term Memory (LSTM) neural network pipeline for predicting depression and anxiety. Transfer learning and hidden Markov models are used in the method to solve the missing observation issue. The purpose of this study is to predict mobility impairment based on WHODAS 2.0 evaluation using digital biomarkers using an attention-based Long Short-Term Memory (LSTM) neural network pipeline. The GAD-7 score was used to validate the pipeline, which performed better than a baseline and correctly predicted the intensity of generalized anxiety based on socio demographic patient data and two wearable/mobile sensor data sets.

In 2023, Casado, et al [20] has introduced a Deep Neural Networks (DNN) for diagnosing depression using physiological signals from facial videos. Using rPPG signals, this method computes more than 60 characteristics, which are then trained into machine learning goes back to identify various depression degrees. It performs better than approaches based on deep learning and manually designed methods. The researchers suggest a unique method that does away with the necessity for contact-based sensors by extracting HRV characteristics from facial footage.

In 2024, Pan, et al [21] has introduced a Depression Recognition Network with Spatial-Temporal Attention (STA-DRN) for depression identification. This method uses an attention-based vector-wise fusion strategy, ResNetstyle modules, and a unique Spatial-Temporal Attention mechanism to improve feature extraction as well as relevancy by collecting both local and global spatial-temporal information. With average absolute error/root mean square error ratings of 6.15/7.98 & 6.00/7.75, respectively, the experimental findings demonstrate

competitive performance.

In 2022, He, et.al [22] has introduced a self-adaptation network (SAN) for depression recognition. Deep learning methods have been developed to measure the degree of depression, a common mental condition. Nevertheless, these algorithms frequently produce noisy labeling because they lack correlations between BDI-II scores and facial images as well as label distribution. ResNet-18 and ResNet-50 for deep feature extraction, self-attention for weights, square ranking regularization for partitions, and re-labeling for dubious annotations make up the architecture's four modules.

Problem statement

Current techniques for estimating the intensity of anxiety and depression based on audio and visual data have poor precision, large error rates, and low accuracy. Using the AVEC2013 and AVEC2014 datasets, we suggest the Adaptive Capsule Network with Electric Eel Foraging Optimization (ACN-EeFO) for enhanced prediction accuracy in order to overcome these problems.

3. PROPOSED METHODOLOGY

The working principle of ACN-EeFO is illustrated in Figure 1. The five processes in the suggested method are: (1) Data collection, (2) Feature extraction, (3) Feature fusion and (4) Classification (5) Optimization.

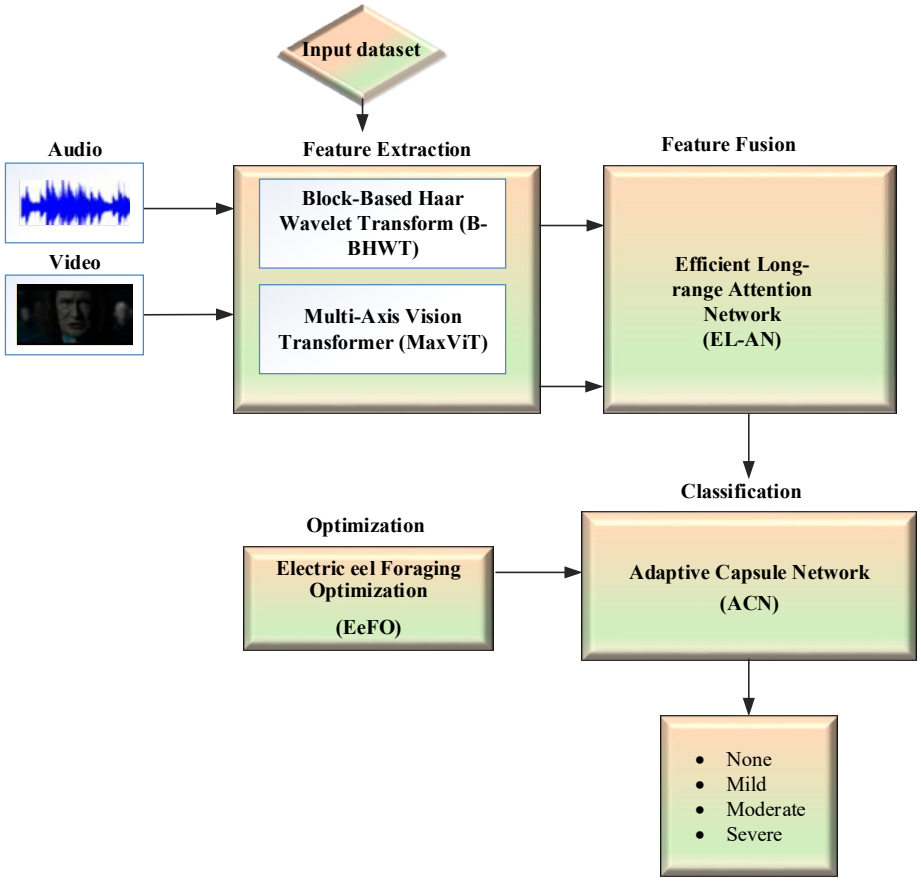

Figure 1: Workflow diagram of proposed ACN-EeFO method

3.1 Data Collection

The proposed model gathers data and conducts research. We obtained all of our data from two datasets, such as AVEC2013 and AVEC2014 .In order to predict the level of depression and anxiety severity this study offers an innovative deep multi-modal framework. Voice clues and facial emotions are the inputs used by the multi-modal network. The audio and video data were first divided into fixed-length frames. After that, two networks are used, one for audio frames and the other for video frames, to extract the spatial features. The following section describes the process of feature extraction of both video and audio frames.

3.2 Feature Extraction

After extracting the frames from the video and audio inputs, feature extraction is carried out. In the feature extraction stage, two networks are used to separately extract features from the extracted video and audio frames.

The Multi-Axis Vision Transformer is used for extracting features from video frames, and the Block-Based Haar Wavelet Transform is used for extracting features from audio frames. The explanation is given below,

3.2.1 Multi-Axis Vision Transformer (MaxViT) Based Feature Extraction of Video

The Multi-Axis Vision Transformer (MaxViT) is leveraged in this study for feature extraction from video data to aid in the multimodal representation of depression and anxiety severity prediction. The Multi-Axis Self-Attention (Max-SA) module introduces an efficient attention mechanism to capture both global and local interactions within the video frames. Block attention and grid attention are the two sparse forms that result from breaking down the conventional completely dense attention process. This method retains the capacity to capture non-local interactions while reducing the complexity of computation from exponential to linear, which makes it more efficient. In block attention, the input feature map (denoted as $Y \in S^{K \times V \times D}$) is partitioned into non-overlapping windows of size $R \times R$. This allows attention mechanisms to focus on smaller, manageable sections of the feature map, thereby reducing computational load and enabling local interactions. Grid attention reshapes the feature map into $Q \times Q$ partitions, allowing for the capturing of global interactions across the entire feature map. This is achieved by reshaping the tensor into a form that can efficiently handle these partitions. The MaxViT model employs a hierarchical design similar to conventional Convolutional Neural Networks (CNNs), with a backbone that includes an initial down sampling stage followed by multiple stages of MaxViT blocks. The MaxViT model is used to extract features from video frames, capturing both local and global patterns within the data.[23-26]

MaxViT is a model that is employed for decoding video frames to extract spatial relationship and structure, as well as global or long-range interactions from videos at varying sizes for feature analysis. This analysis of multiple states allows for recognizing common characteristics and differences depending on the patient's emotional condition. The hierarchical design integrates features at various layers of abstraction, which increases the model's predictive capabilities when it comes to the severity of depression and anxiety.

3.2.2 Block-Based Haar Wavelet Transform (B-BHWT) Based Feature Extraction of Audio

The Block based Haar wavelet transform known as B-BHWT is very useful in signal processing especially in feature extraction of one dimensional signal like audio frame. Those multi-modal data can be of use in instances such as estimating, the severity of depression and anxiety. The HWT is a form of wavelet transform that uses Haar wavelets that are among the simplest wavelets. It breaks down a signal into a number of wavelet coefficients, whereby the information on the time and the frequency is obtained. This decomposition is obtained by using the Haar wavelet basis functions on the signal provided. The one-dimensional audio signal is divided into H blocks, denoted as Aj where j =1, 2, 3,…, H. Each block Aj is of size $1 \times M$ .The Haar matrix B is defined using Haar wavelet basis functions. The forward HWT for every block can be calculated using the given equation (1):

$$L = AB^{L} \tag{1}$$

where, A is the signal block and the B denotes the Haar matrix. L represents the transform coefficient matrix .The Haar wavelet matrix B of size $1 \times M$ is constructed using the formulas for the basic functions $B_j(y)$ it's given by equation (2 and 3):

$$B_{(0)}(y) = \frac{1}{\sqrt{M}}, \qquad (0 \le y \le Y) \tag{2}$$

$$B_{(1)}(y) = \frac{1}{\sqrt{M}} \begin{cases} 1, & 0 \le y \le Y \\ -1, & \frac{y}{2} \le y \le Y \\ 0, & otherwise, \end{cases} \tag{3}$$

These basis functions form the rows of the Haar matrix B .The kernel matrix for the HWT is generated from the Haar basis functions. For example, for $M = 8$, the 8×8 Haar wavelet matrix is explicitly given, showing the structure of the transform. The original signal block can be reconstructed from the transform coefficients using the inverse Haar transform it's given by equation (4):

$$BP = LB \tag{4}$$

where, P *and* L is the reconstructed signal block, and B is the transpose of the Haar matrix.

The audio signal is divided into frames, and each frame is further divided into smaller blocks as described above. The HWT is applied to each block of the audio frames to extract the wavelet coefficients. These coefficients represent the signal's characteristics in both time and frequency domains. Features are selected from the wavelet coefficients that are most relevant to the classification or regression task at hand (predicting depression and anxiety severity).The extracted features from audio are combined with features from other modalities (e.g., text, video) to create a comprehensive feature set. Machine learning models are trained using these multimodal features to predict the severity of depression and anxiety.

The Block-Based Haar Wavelet Transform (HWT) is the technique of selective feature extraction on audio signals for machine learning predictions of depression and anxiety severity. This process helps in converting the raw data into discrete features and then combined to be further analyzed.

3.3 Efficient Long-range Attention Network (EL-AN) based Feature Fusion

The Efficient Long-range Attention Network (EL-AN) is designed to enhance the performance of multimodal representation in tasks such as depression and anxiety severity prediction. This is achieved by leveraging both audio and video features. The overall pipeline of EL-AN, which is focused on image super-resolution in the provided context, can be adapted for multimodal fusion. Below is an explanation of how EL-AN could be applied to audio-video feature fusion for predicting depression and anxiety severity. The deep features from both modalities are combined to predict the severity of depression and anxiety. The deep features from audio and video are combined, possibly through concatenation or addition. Y_k is given by equation (5):

$$Y_k = K_{PD}(Y_r + Y_f) \tag{5}$$

where, K_{PD} denotes the reconstruction module. Y_r *and* Y_f are the inputs. Y_k is the fused features representation. A reconstruction or prediction module processes the fused features to predict the severity of depression and anxiety. This could involve a simple neural network layer or a more complex mechanism depending on the application. The EL-AN pipeline adapted for audio-video feature fusion in depression and anxiety severity prediction involves fusing the deep features to form a comprehensive representation. This approach allows for efficient and effective multimodal representation, crucial for tasks involving complex emotional and psychological assessments. Then the fused features are given to the classification.

3.4 Adaptive Capsule Network (ACN) Based Classification

The fused features are given to the ACN for classifying the Multimodal representation of depression and anxiety severity. The Adaptive Capsule Network (ACN) presents an advanced approach for classifying and predicting depression and anxiety severity using multimodal representations. The detailed explanation of how it operates is given below,

- **Adaptive Capsule Layer**

A convolution operation, denoted as E, is applied to these feature maps Y^0 , generating an intermediate representation is given by equation (6):

$$Y^0(d, M^t \times C^t \times h \times h, t) \tag{6}$$

where, M^t represents the number of primary capsules, C^t is the capsule dimension, h the filter size, and t the local receptive area. The intermediate representation is reshaped to form capsule vectors is given by equation (7):

$$Y^t_{(d, M^t \times C^t \times h \times h.t)} = E(Y^0_{(d, e.k, v)}), \tag{7}$$

where, $n^t = h \times h \times M^t$ is the number of capsules and C^t is the capsule dimension. A learnable fixed spatial bias tensor $R^t_{(t, n^t, C^t)}$ initialized with a Normal Distribution (0, 1) is added to these capsule vectors. This tensor preserves spatial information and encodes spatial relationships among capsule vectors. The capsule vectors are adjusted by adding the expanded spatial bias tensor, resulting in given by equation (8):

$$\vec{Y}_{(d, t, n^t, C^t)} = Y^t_{(d, t, n^t, C^t)} + R^t_{((1, t, n^t, C^t))} \tag{8}$$

Capsule vectors are processed through a small network which includes RMSNORM regularization, two fully connected layers L^1 *and* L^2 , and activation functions ReLU and Sigmoid. The RMSNORM operation standardizes the capsule vectors to enhance model convergence. This small network outputs the local adaptive value tensor $B^t_{(d,n^t,C^t)}$ .Local adaptive values are gathered to form global adaptive values $B^f_{(d,t,n^t,C^t)}$ .These global adaptive values are added to the capsule vectors $Y^t_{(d,t,n^t,C^t)}$ to produce the final adaptive capsule vectors is given by equation (9):

$$X^t_{(d,t,n^t,C^t)} = B^f_{(d,t,n^t,C^t)} + Y^t_{((d,t,n^t,C^t))} \quad (9)$$

- **Full Convolutional Capsule Layer**

The full convolutional capsule layer utilizes two learnable weight matrices: pose weight matrices V^{pose} and route weight matrices V^{route} .Routing coefficients $d_{a,b}$ and prediction vectors $v_{b|a}$ are generated based on these matrices. The existence probabilities of capsule vectors, i^t_a and i^{t+1}_b , in different capsule layers are calculated. A weighted average operation on prediction vectors generates high-level capsule vectors are given by equation (10):

$$v_{b|a} = V^{pose}_{a,b} X^t_a \quad (10)$$

The ACN is designed to integrate and analyze multimodal data effectively, which is crucial for accurately predicting depression and anxiety severity. The Adaptive Capsule Network utilises convolutional layers and capsule layers, with spatial biases that are optimised during learning, to create the final model for predicting the severities of depression and anxiety. It can handle different kinds of data and is beneficial when it comes to predicting the excerpt of the representation. In order to enhance the accuracy and minimize the errors and their frequency in practical implementations, the network is refined with help of EeFO, which, in turn, leads to decrease in the time and amount of computation and its cost.

3.5 Electric eel Foraging (EeFO) Based Optimization

The Electric eel Foraging Optimization (EeFO) is used to optimize the learning parameters of ACN. Electric Eel Foraging Optimization is an algorithm inspired by the hunting and social behaviors of electric eels. The EeFO mimics this cooperative hunting strategy to solve optimization problems. The conversing, resting, traveling, and hunting activities of electric eels serve as an inspiration for the EeFO exploration and exploitation stages.

Step1: Initialization

Create an initial population of Electric eel Foraging Optimization solutions, each representing a set of ACN hyper parameters.

Step 2: Generation of Random Variables

Generate at random the optimization variables of Electric eel Foraging Optimization to attain the best solution.

Step 3: Evaluation of Fitness Function

The role of fitness function is used to forecast the desired outcome, which is to appropriately categorize by positive and negative zones. The fitness function equation is given by equation (11):

$$fit(y_j(l)) \le fit(u_j(l+1)) \quad (11)$$

where , $fit(y_j)$ denotes the fitness of the candidate position of the j -th electric eel.

Step 4: Migrating (Exploration) for improving accuracy

Electric eels move from their resting area to the hunting area to find prey. The movement is described by the equation (12):

$$u_j(l+1) = -s_5 \times S_j(l+1) + s_6 \times K_s(l+1) - T \times (K_s(l+1) - y_j(l)) \quad (12)$$

where, K_s denotes the any position within the hunting area, s_5 *and* s_6 represents a random numbers within $(0,1)$. $(K_s(l+1) - y_j(l))$ denotes the indicates that eels move towards the hunting area. T represents the

levy flight function. Eels explore new positions by combining their current position with random factors and a hunting area position. Introduces randomness to avoid local optima and explore the search space more effectively. Ensures eels move to positions with better fitness values, balancing exploration and exploitation.

Step 5: Hunting (Exploitation) for reducing error rate, processing time, computational complexity and cost

The exploitation phase in this context refers to the process where the algorithm refines its search around promising solutions, focusing on a smaller, more specific region to find the optimal solution. Initially, the hunting area is defined around the prey using a parameter γ_0. This area is determined by the distance between the eel and the prey and scales as time progresses. In the exploitation phase, the eel performs a curling movement where it wraps around the prey. The hunting area can be given by equation (13):

$$\{Y \left| Y - y_{prey}(l) \right| \le \gamma_0 \times \left| \bar{y}(l) - y_{pery}(l) \right| \} \quad (13)$$

where, γ_0 denotes the initial scale of hunting area. y_{prey} is the eel centers with its hunting range determined by the $\gamma_0 \times \left| \bar{y}(l) - y_{pery}(l) \right|$. Initially, the algorithm explores a wider area to identify potential prey (solutions). As it progresses, it shifts to exploitation by narrowing the search area and focusing on refining solutions within this smaller region. The positions visited by the prey (noted as red dots) are used to update the eel's position. This iterative updating helps in honing in on the optimal solution. In EeFO, the exploitation phase is characterized by reducing the hunting area size and focusing the search on a smaller, more promising region.

Step 6: Termination

Once the best answers are obtained using equations (12), Additionally, equation (13) yields the most accurate answer, and minimizes error rates, processing times, computing complexity, and cost. This iteration is remaining until the tentative criteria $j = j + 1$ is met. Finally, the suggested ACN- EeFO method accurately categorizes the multimodal representation of depression and anxiety severity.

Hence, ACN detailed and explained. It Input Video and Audio frames features are separately extracted using MaxViT and B-BHWT methods, Features are fused using EeFO method and Classification using ACN optimization using EeFO method for multimodal representation of depression and anxiety severity prediction demonstrating superior efficiency and accuracy. In the next section the results and discussions are discussed.

4. RESULT AND DISCUSSIONS

This section describes the introduced scheme's findings and debate. Python is used to carry out the Result and discussion of the method. Here is some of the Implementation parameter is mentioned in table 1:

Table 1: Implementation Parameters

Parameters	Description
Proposed Neural Network	**ACN-EeFO**
OS	**Windows 10**
Optimization	**EeFO**
Datasets	**AVEC2013, AVEC2014 datasets**
Software	**Python 3.7**

4.1 Dataset Description

For predicting depression and anxiety, two datasets are taken; namely AVEC2013, AVEC2014 datasets and its descriptions are given below:

4.1.1 AVEC2013 dataset

The AVEC2013 subset of the larger AViD-Corpus is made up of 340 video clips that were captured by 292 people using webcams and microphones while they were engaged in human–computer interaction (HCI) tasks. In order to complete HCI activities, participants, ages 18 to 63, recorded audio and video clips at 41 kHz, 30 frames per second, and 640 x 480 pixels. There were 50 recordings total, split into test, development, and training sets using the H.264 code. The BDI-II score is used in the AVEC2013 dataset respectively.

4.1.1. AVEC2014 dataset

The dataset includes 150 movies for two tasks: North wind, where participants read a fragment from a tale, and Freeform, where participants discuss a painful childhood memory. Every participant speaks German, and they were all captured on camera and microphone. The dataset is split up into test, development, and training sets. A

BDI-II score is assigned to every movie. The corpus includes training, development, and test sets with 50 recordings each, totaling 300 movies with 84 participants who are at least 31.5 years old. Table 2 shows the Comparison for performance analysis of proposed approach. Table 2 shows the BDI-II scores in both AVEC2013 and AVEC2014 datasets,

Table 2: BDI-II scores in both AVEC2013 and AVEC2014 datasets

BDI-II score	Depression Stages
0-16	None
17-25	Mild
27-45	Moderate
50-65	Severe

The table 2 categorizes BDI-II (Beck Depression Inventory-II) scores into different depression stages for the AVEC2013 and AVEC2014 datasets. Scores of 0-16 indicate no depression, 17-25 indicate mild depression, 27-45 indicate moderate depression, and 50-65 indicate severe depression, providing a framework for assessing depression severity. Table 3 shows the performance comparison of AVEC2013 dataset.

Table 3: Performance comparison of AVEC2013 dataset

Methods / Performance	LSTM[8]	DNN[9]	STA-DRN[10]	SAN[11]	ACN-EeFO (Proposed)
Accuracy (%)	98.56	97.45	96.45	98.34	99.82
Precision (%)	94.45	93.57	92.89	96.44	99.72
Specificity (%)	87.56	89.33	96.56	92.90	98.58
Sensitivity (%)	89.56	78.67	86.90	76.89	99.62
F1-score (%)	92.44	94.55	97.67	87.90	99.23
Error rate (%)	0.5	0.3	0.2	0.4	0.1

The table 3 compares the performance of various methods on the AVEC2013 dataset. The proposed ACN-EeFO method outperforms others across all metrics: accuracy (99.82%), precision (99.72%), specificity (98.58%), sensitivity (99.62%), F1-score (99.23%), and error rate (0.1%). It shows significant improvements, particularly in accuracy, precision, and sensitivity. Table 4 shows the Performance comparison of AVEC2014 dataset.

Table 4: Performance comparison of AVEC2014 dataset

Methods / Performance	LSTM[8]	DNN[9]	STA-DRN[10]	SAN[11]	ACN-EeFOA (Proposed)
Accuracy (%)	93.58	93.84	96.67	98.36	99.87
Precision (%)	92.53	83.57	82.89	86.44	99.75
Specificity (%)	87.62	89.33	96.56	82.90	98.63
Sensitivity (%)	81.62	77.67	81.90	78.89	99.76
F1-score (%)	92.40	94.64	97.67	87.90	99.81
Error rate (%)	0.4	0.4	0.3	0.5	0.2

The table 4 compares various methods' performance on the AVEC2014 dataset. The proposed ACN-EeFOA method achieves the highest accuracy (99.87%), precision (99.75%), specificity (98.63%), sensitivity (99.76%), F1-score (99.81%), and lowest error rate (0.2%), outperforming other techniques significantly across all metrics. Table 5 shows the Comparison of Error Rates for AVEC2013 and AVEC2014 datasets

Table 5: Comparison of Error Rates for AVEC2013 and AVEC2014 datasets

Dataset	Methods	MAE (%)	RMSE (%)
AVEC2013	LSTM[8]	2.45	3.56
	DNN[9]	3.34	4.34
	STA-DRN[10]	4.23	5.34
	SAN[11]	7.10	8.23
	ACN-EeFOA(Proposed)	9.34	9.56
AVEC2014	LSTM[8]	4.23	5.34
	DNN[9]	6.34	7.45
	STA-DRN[10]	9.34	9.57

	SAN[11]	8.25	9.48
	ACN-EeFOA(Proposed)	9.82	9.72

The Table 5 compares the error rates of various methods on the AVEC2013 and AVEC2014 datasets. Metrics include Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The proposed method, ACN-EeFOA, shows higher error rates compared to other methods, indicating room for improvement in minimizing prediction errors. The figure 2 shows the Confusion matrix for (a) AVEC2013, (b) AVEC2014 datasets,

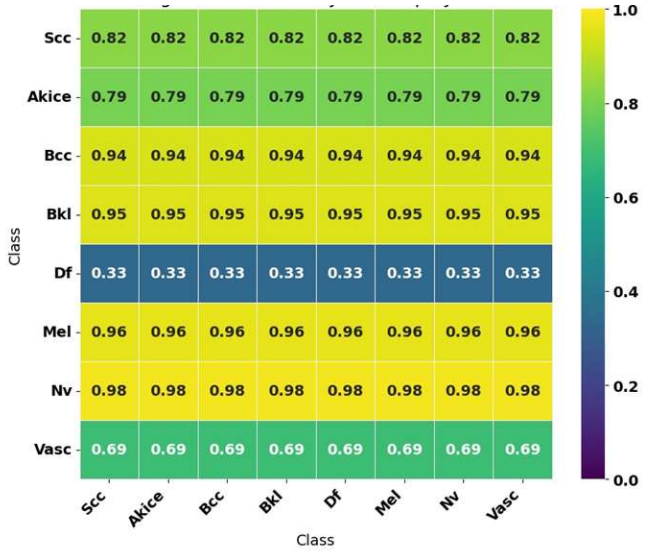

Figure 2: Confusion matrix for (a) AVEC2013, (b) AVEC2014 datasets

The Figure 2 shows the confusion matrix's performance analysis for (a) AVEC2013, (b) AVEC2014 datasets .The suggested method performs exceptionally well in identifying real instances and minimizing errors across all datasets, as confirmed by the confusion matrix analysis. This leads to high accuracy, precision, recall, and overall robust performance in SC detection. Figure 3 shows the estimated and ground truth values for depression and anxiety severity prediction,

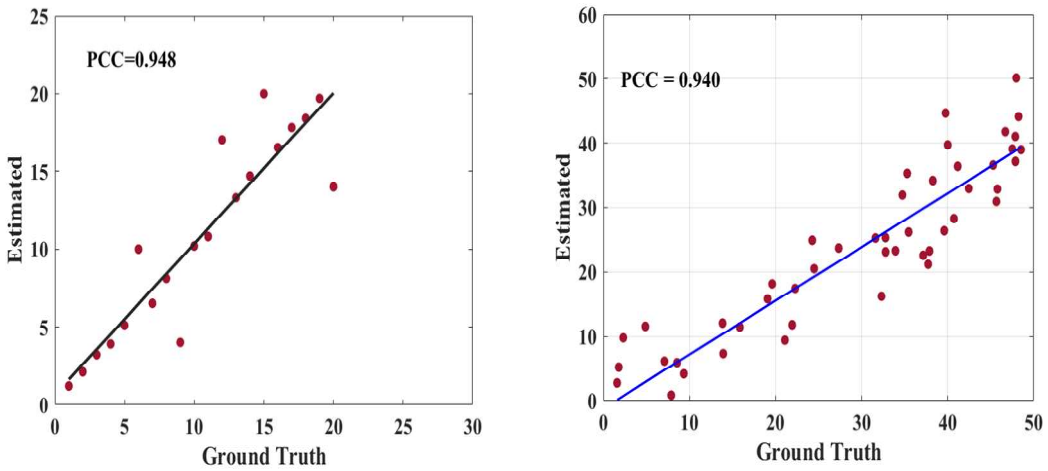

Figure 3: Estimated and ground truth values for depression and anxiety severity prediction

The Figure 3 shows the estimated and ground truth values for depression and anxiety severity prediction. Plot (a) for one dataset exhibits a high correlation (PCC = 0.948), whereas plot (b) for another dataset reveals a little lower but still substantial correlation (PCC = 0.940).

5. CONCLUSION

In this manuscript ACN-EeFO is successfully manipulated, the input data is taken from two dataset such as AVEC2013 and AVEC2014. Following that, the features of audio and video are extracted using MaxViT and B-

BHWT. Then, the extracted audio and video data are fused using Efficient Long-range Attention Network (EL-AN). After that the classification and optimization are done using ACN-EeFO for classify the different stages of depression levels using BDI-II score. The introduced system is executed in python. The efficiency of the proposed ACN-EeFO is analyzed using 2 datasets and attains 99.82% accuracy and 0.1% error rate, compared with the existing methods. In the future, instead of merging the audio and video features at the end, it will attempt to find out if doing so enhances prediction performance. Additionally, scientists plan to utilize this framework for additional tasks including the prediction of various diseases.

REFERENCES:

1. Zhou, Enqi, et al. "Prediction of anxious depression using multimodal neuroimaging and machine learning." NeuroImage 285 (2024): 120499.
2. Habets, Philippe C., et al. "Multimodal data integration advances longitudinal prediction of the naturalistic course of depression and reveals a multimodal signature of remission during 2-year follow-up." Biological psychiatry 94.12 (2023): 948-958.
3. M. Preetha, Archana A B, K. Ragavan, T. Kalaichelvi, M. Venkatesan "A Preliminary Analysis by using FCGA for Developing Low Power Neural Network Controller Autonomous Mobile Robot Navigation", International Journal of Intelligent Systems and Applications in Engineering (IJISAE), ISSN:2147-6799. Vol:12, issue 9s, Page No:39-42, 2024.
4. Ahmed, Sabbir, et al. "Taking all the factors we need: A multimodal depression classification with uncertainty approximation." IEEE Access (2023).
5. M. Preetha, Raja Rao Budaraju, Jackulin. C, P. S. G. Aruna Sri, T. Padmapriya "Deep Learning-Driven Real-Time Multimodal Healthcare Data Synthesis", International Journal of Intelligent Systems and Applications in Engineering (IJISAE), ISSN:2147-6799, Vol.12, Issue 5, page No:360-369, 2024
6. A. Nithya, M. Raja, D. Latha, M. Preetha, M. Karthikeyan and G. S. Uthayakumar, "Artificial Intelligence on Mobile Multimedia Networks for Call Admission Control Systems," 2023 4th International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2023, pp. 1678-1683, doi: 10.1109/ICOSEC58147.2023.10275999.
7. Balaji Singaram, M.S.Vinmathi, Dr.H.B.Michael Rajan, Jeyamohan H, T. Manikandan, "Data-Driven Estimation of Lithium-Ion Battery State-of-Health Prediction Approach Using Machine Learning Algorithm for Enhanced Battery Management Systems", Nanotechnology Perceptions, ISSN 1660-6795 2024, Vol: 20, 7s, 93-103
8. K Siva Kumar, Dr. Raghu Dhumpati , Dr. R. Benschwartz , Gowtham A , Dr. A. Jeeva , Dr. S. Devikala " A Fuzzy Logic Approach for Efficient Control of BLDC Motors for Low-Cost Electric Two-Wheelers " Nanotechnology Perceptions, https://doi.org/10.62441/nano-ntp.v20iS5.16 ISSN 1660-6795 2024, Vol: 20, 5s, 191-202.
9. Dr.M.Preetha, Balaji Singaram, Dr.I. Manju, B.Hemalatha, P. Bhuvaneswari "Machine Learning in Breast Cancer Treatment for Enhanced Outcomes with Regional Inductive Moderate Hyperthermia and Neoadjuvant Chemotherapy" Nanotechnology Perceptions, ISSN 1660-6795 2024, Vol: 20, 5s, 245-259
10. Cohen, Joshua, et al. "A multimodal dialog approach to mental state characterization in clinically depressed, anxious, and suicidal populations." Frontiers in psychology 14 (2023): 1135469.
11. Balaji Singaram, Lakshmi. B, Dr.M.Preetha, V.K. RamyaBharathi, Dr.S.Muthumarilakshmi, Rakesh Kumar Giri "A Smart IoT-Based Fire Detection and Machine Learning Based Control System for Advancing Fire Safety" Nanotechnology Perceptions, ISSN 1660-6795 2024, Vol: 20, 5s, 229-244.
12. K Sivakumar, A. Saravanan, Vijaya Saraswathi R, T. Mahalingam, S. Devikala, T.R. Ramesh "Beyond the Current State of the Art in Electric Vehicle Technology in Robotics and Automation" Journal of Electrical Systems, https://doi.org/10.52783/jes.1254 ISSN 1112-5209 2024, Vol: 20, 4s, 2282-2291.
13. S. Devikala, Rabi.J, V.P.Murugan, J.S. Christy Mano Raj, K. Mohanasundaram, K Sivakumar, "Development of fuzzy logic controller in Automatic Vehicle Navigation using IOT.," Journal of Electrical Systems, https://doi.org/10.52783/jes.1254 ISSN 1112-5209 2024, Vol: 20, 3s, 114-121.
14. Yoo, Joo Hun, et al. "Mood Disorder Severity and Subtype Classification Using Multimodal Deep Neural Network Models." Sensors 24.2 (2024): 715.
15. Zhang, Zhenwei, et al. "Multimodal Sensing for Depression Risk Detection: Integrating Audio, Video, and Text Data." *Sensors* 24.12 (2024): 3714.

16. M. Mohammed Thaha, M. Preetha, K Sivakumar, & Rajendrakumar Ramadass " An Aerial Robotics Investigation into the Stability, Coordination, and Movement of Strategies for Directing Swarm and Formation of Autonomous MAVs and Diverse Groups of Driverless Vehicles (UGVs)," International Journal on Recent and Innovation Trends in Computing and Communication https://doi.org/10.17762/ijritcc.v11i3.8908 ISSN: 2321-8169 Volume: 11 Issue: 3 ,February 2023.
17. Sükei, Emese, et al. "Automatic patient functionality assessment from multimodal data using deep learning techniques–Development and feasibility evaluation." Internet Interventions 33 (2023): 100657.
18. K Sivakumar, J.V Sai Prasanna Kumar, K Loganathan, V Mugendiran, T Maridurai, K Suresh "Machining characteristics of silane-treated wheat husk biosilica in deionized water dielectric on EDM drilling of Ti-6Al-4 V alloy," Biomass Conversion and Biorefinery Published online: 1 February 2022 https://doi.org/10.1007/s13399-022-02308-4. ISSN 2190-6823.
19. Thati, Ravi Prasad, et al. "A novel multi-modal depression detection approach based on mobile crowd sensing and task-based mechanisms." Multimedia Tools and Applications 82.4 (2023): 4787-4820. ….AVEC2013 and AVEC2014
20. Casado, Constantino Álvarez, Manuel Lage Cañellas, and Miguel Bordallo López. "Depression recognition using remote photoplethysmography from facial videos." *IEEE Transactions on Affective Computing* 14.4 (2023): 3305-3316
21. Pan, Yuchen, et al. "Spatial–temporal attention network for depression recognition from facial videos." Expert Systems with Applications 237 (2024): 121410.
22. He, Lang, et al. "Reducing noisy annotations for depression estimation from facial images." Neural Networks 153 (2022): 120-129.
23. Ong, Kah Liang, et al. "SCQT-MaxViT: Speech Emotion Recognition With Constant-Q Transform and Multi-Axis Vision Transformer." *IEEE Access* 11 (2023): 63081-63091.
24. Khafaga, Doaa Sami, et al. "Compression of Bio-Signals Using Block-Based Haar Wavelet Transform and COVIDOA for IoMT Systems." *Bioengineering* 10.4 (2023): 406.
25. Zhang, Xindong, et al. "Efficient long-range attention network for image super-resolution." *European conference on computer vision*. Cham: Springer Nature Switzerland, 2022.
26. Tao, Jianwei, et al. "Adaptive capsule network." *Computer Vision and Image Understanding* 218 (2022): 103405.