Multi-model Centric Facial Expression Recognition Using Improved Coati Optimization-based Exponential Linear Gated Recurrent Unit with Layer Normalization

Laxmi[a,*], Lakshmi Patil[b]

[a, *]Research Scholar and Assistant Professor, Department of Electronics and Communication Engineering, Sharnbasva University, Kalaburagi-585103, Karnataka, India.

[b] Professor and Dean, Department of Electronics and Communication Engineering, Sharnbasva University, Kalaburagi-585103, Karnataka, India.

*Corresponding Author E-mail : laxminijanand@gmail.com

ORCID: 0009-0008-0182-5684

Abstract

Human facial expressions and emotions play vital role in day-to-day communication, and recognizing them is one of the significant tasks in Human-Computer Interfaces (HCI) field. Facial Expression Recognition (FER) utilizing images involves analyzing facial features to recognize and interpret emotions accurately. However, lightning issues, posture changes, and occlusion significantly impact accurate recognition of facial expressions. In this research, an Improved Coati Optimization Algorithm-based Exponential Linear Unit-Gated Recurrent Unit (ICOA-ELU-GRU) with layer normalization is proposed to accurately recognize FER. COA is improved using search and encirclement strategy and lens-opposition learning strategy which improves convergence speed, avoids local optima, and increases global exploration. Images are obtained from affect Net and CK+ datasets, and those images are pre-processed using min-max normalization which improves image quality. Then, VGG16, ResNet 18, and AlexNet are employed to extract features and ICOA selects valid features from extracted features that enhance model efficiency. Finally, ELU-GRU with layer normalization is performed to recognize FER by capturing spatial and temporal data. When compared to existing techniques like Adaptive Multi-layer Perceptual Attention Network (AMP-Net), and multi-attention module, the proposed approach achieves better accuracy of 77.67% and 97.36% for both datasets.

Keywords: Exponential Linear Unit, Facial expressions, Improved Coati Optimization Algorithm, Physical exertion-based escape strategy, and Search and Encirclement strategy

1. Introduction

Emotion recognition is a significant way to understand the behavior of human emotions and generates communication among humans and machines [1]. The recognition of human emotion is executed in numerous ways including facial expression, speech data, physiological parameters, body gestures, and so on [2]. Recent advancements in computer vision highly provide the development of sophisticated effective computing systems in Facial Expression Recognition (FER) [3]. FER's main goal is to recognize the disparate facial expressions of humans from their facial images [4] [5]. The representation of emotion is split into two models, dimensional and discrete models. In dimensional technique, emotions are mapped into arousal,

valence, and dominance dimensions [6]. Human emotion recognition via image sequences enables to define a real and continuous process that assists in obtaining a higher veracity to feeling found [7]. Initially, research on emotion recognition systems depended on a single modality; the most common is the recognition of facial expressions. The technology utilizes biometric techniques to map, analyze, and confirm the identity of faces on video or images [8]. FER has receiving enhance attention in different research domains like road safety, driving behavior, mobile internet areas, Human-Computer Interaction (HCI) [9], computer vision techniques, and behavior of human understanding in multimedia data [10] [11].

FER has receiving enhance attention in different research domains like road safety, driving behavior, mobile internet areas, Human-Computer Interaction (HCI) [9], computer vision techniques, and behavior of human understanding in multimedia data [10] [11]. The procedure of recognizing emotions is dynamic and it focuses on an individual's state of emotion; hence, every person has a specific set of feelings which is associated with their actions [12]. Differences among facial expressions are commonly dependent on facial muscles of subtle distortion [13]. Moreover, it is influenced by non-subjective factors like pose and lightning variations [14]. Researchers have categorized emotions into eight basic classes: neutral, disgust, happiness, anger, sadness, contempt, fear, and surprise which are known as primary emotions. Analyzing facial features and movements makes the recognition of human emotions with significant accuracy [15]. However, lightning issues, posture changes, and occlusion significantly impact the accurate recognition of facial expressions. Hence, ICOA-ELU-GRU with layer normalization is proposed to accurately recognize facial expressions. GRU captures both spatial and temporal data effectively. ELU helps to solve the vanishing gradient issue and layer normalization normalizes the layer parameter which maintains a stable distribution. By performing these processes, the proposed approach achieves accurate recognition in FER.

The primary contribution of this research is as follows:

- VGG16, ResNet 18, and AlexNet are used to extract the features of spatial data, high-level and low-level features that capture meaningful patterns, variations, and discriminative features effectively.
- COA is improved by using a search and encirclement approach by transmitting sound which achieves convergence speed and enhances the effectiveness in FER. Physical exertion-based escape approach is used for escaping while being chased which helps to not fall from local optima and increase the model's adaptability in facial recognition tasks. A lens-opposition learning approach is employed to explore a greater range of facial expressions that increases global exploration capability and makes more understanding of different emotional states.
- ELM-GRU with layer normalization is used to recognize the facial expressions. GRU captures both spatial and temporal dependency, ELU solves vanishing gradient issues, and layer normalization helps to maintain stable distribution. By performing this process, the proposed technique achieves accurate recognition in FER.

The remaining portion is organized as: Section 2 indicates the literature survey. Section 3 explains about proposed methodology. Section 4 discusses the results. Section 5 contains the overall conclusion of the paper.

2. Literature Survey

The related work about facial expression recognition was discussed below briefly along with their merits and demerits.

Hanwei Liu *et al.* [16] implemented an Adaptive Multi-layer Perceptual Attention Network (AMP-Net) to recognize facial expressions effectively. AMP-Net extracts local, global, and salient features of facial emotion with various fine-grained features to learn key data and the underlying diversity of facial emotions. A global perception approach was established to acquire features with various receptive fields, and an attention perception technique increase salient features of emotions depending on prior knowledge. This approach effectively removes invalid data under occlusion by dynamically adjusting attention to appropriate facial features and producing high robustness to variant poses and facial occlusion. However, this approach suffers from overfitting issues because of limited training data which hinders the capability to generalize well on unseen facial expressions.

Yuanlun Xie *et al.* [17] developed a Feature Extraction Module (FEM)-based on three dense blocks to recognize facial expressions that extract multi-level features as an outcome of an entire network of feature extraction. A Feature Fusion Module (FFM) was utilized with local and global attention to fuse these various feature levels in pairs in a top-down way and a new facial feature expression was constructed. This approach enhances the performance of traditional convolutional backbone networks on recognizing facial expressions to a certain extent. However, FEM has low recognition accuracy in certain facial expressions like disgust and fear due to subtle and nuanced facial features which leads to challenges in detecting and classifying accurately.

Bei Pan *et al.* [18] introduced a multimodal fusion framework depending on a Genetic Algorithm (GA) and Extreme Learning Machine (ELM) to recognize emotions in video clips. To solve the visual modality, keyframes were initially extracted from a sequence of consecutive images and then geometric feature representation was defined for detecting keyframe transformation. Then, relevant features were chosen using evolutionary optimization to minimize feature dimensions. Finally, an optimized ELM classifier was established to recognize unimodal emotions. The model structure and optimized emotional features not only significantly enhance the emotion recognition accuracy in audio and video modalities but also minimize model complexity. However, GA-ELM lacks suboptimal performance because of GA's exploration limitations and ELM's potential challenges in capturing intricate relationships with multimodal data.

Junnan Zhi *et al.* [19] suggested a multi-attention module to recognize dynamic facial emotion. Depending on ResNet, the approach employs attention modules in three dimensions of channel, space, and time to choose the video parts which was significant for recognizing specific emotions. This module computes the channel and spatial attention on feature map of every frame and weights them. At the end of extracted feature network, the frame attention was computed depending on data included in feature vector for every output frame. This approach improves the feature extraction ability of Convolutional Neural Network (CNN) by constructing various attention techniques. However, the accuracy of recognition was not satisfactory for small-scale locally collected data due to inconsistent distribution data.

Moutan Mukhopadhyay *et al.* [20] presented textural image features like Local Ternary Pattern (LTP), Local Binary Pattern (LBP), and Completed LBP (CLBP) based on CNN to recognize facial expressions. This approach employs the benefits of textual features that were greatly correlated with the changes in facial expression and thereby trains a CNN technique to recognize the expression. The textual image has less low-level data than grey-scale image, therefore, when CNN model training was generated with these images, then the approach learns higher with normal images which have a high level of data. Therefore, this approach achieved enhanced emotion recognition accuracy with the help of CNN. However, this approach has limited capability to capture subtle variations in facial expression due to its main focus on capturing local texture patterns.

In overall evaluation, the existing methods have limitations like low recognition accuracy in certain facial expressions, overfitting issues because of limited training data, and challenges in capturing complex relationships with multimodal data. To overcome this issue, the ELM-GRU with layer normalization is proposed to accurately recognize facial expressions which enhances the recognition accuracy.

3. Proposed Methodology

In this research, the ELM-GRU with layer normalization is proposed to accurately recognize facial expressions. The image is obtained from two benchmark datasets namely AffectNet and CK+. The min-max normalization is used for pre-processing which improves image quality by performing scaling. Then, features are extracted using VGG16, ResNet18, and AlexNet and then ICOA is employed to select the valid features from extracted ones. Finally, ELM-GRU with layer normalization is used for FER. Figure 1 indicates a block diagram for proposed approach.

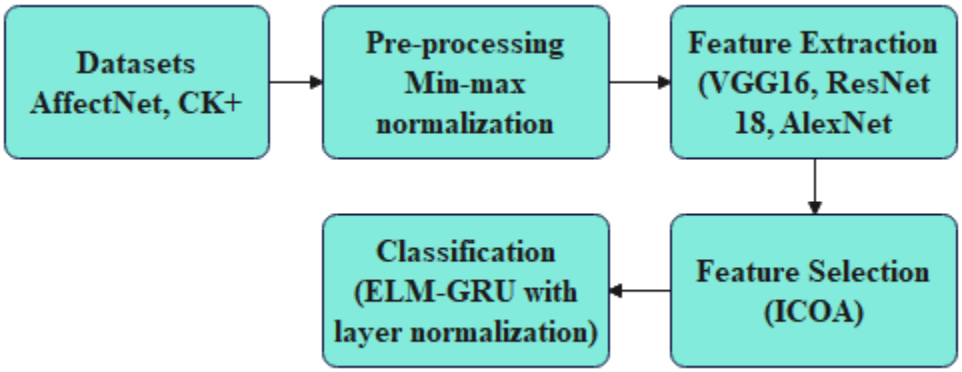

Figure 1. Block diagram for proposed approach

3.1 Datasets

In this research, AffectNet [21] and CK+ [22] datasets are used to evaluate the proposed approach. These datasets are widely utilized to recognize facial expressions that generate diverse facial expressions in different contexts.

3.1 AffectNet

AffectNet is the largest dataset in facial expression gathered from Kaggle. In this dataset, eight basic expressions like neutral, fear, surprise, happiness, anger, disgust, sadness, and contempt are utilized and it has a 96 x 96-pixel image. It contains 4000 images for testing, 28, 7652 images for training, and 4000 images for validating purposes respectively. Figure 2 indicates examples for AffectNet dataset.

Neutral Fear Surprise Happy Anger Disgust Sadness Contempt

Figure 2 Examples of AffectNet datasets

3.2 CK+ dataset

The CK+ dataset has 593 sequence videos from 123 various people, a person's ages ranging from 18 to 50 years with different genders respectively. It has image frames with either 640 x 480 or 640 x 490-pixel resolutions. It contains seven basic expressions without neutral expressions. Figure 3 represents examples of the CK+ dataset. These obtained images are fed into pre-processing phase for normalization.

Fear Surprise Happy Anger Disgust Sadness Disgust

Figure 3 Examples of CK+ dataset

3.2 Pre-processing

After obtaining images, the min-max normalization [23] is utilized in a pre-processing stage which focuses on improving the image quality. This stage involves scaling the pixel values of images to a standard range of 0 to 1 to make consistent data input for the model which is efficient by using min-max normalization. It makes every input data on a similar scale which prevents some features from dominating the learning process. By scaling the image from 0 to 1, the color value will be adjusted which enhances the image quality. Additionally, minor noises are removed in this stage. The min-max normalization processes every intensity of a pixel which is formulated in equation (1)

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

Where x_i determines value of a pixel intensity in image, $\max(x)$ and $\min(x)$ illustrates maximum and minimum values for image pixel intensity. The x_{norm} represents normalized pixel intensity respectively. By using these equations, it effectively increases the image quality by

scaling the images. This enhanced image quality is fed as input to the feature extraction process through the neural network's input layer.

3.3 Feature Extraction

After obtaining enhanced image quality, the VGG16, ResNet18, and AlexNet are utilized to extract the features in FER. Transfer learning allows leveraging knowledge from one task (ImageNet classification) to another (FER). The deep layers of these approaches capture hierarchical features that assist in recognizing complex patterns in facial expression. These approaches load pre-trained model weights, remove the classification head (Fully Connected layers (FC)) from the model, and then features are extracted from remaining layers of model. A detailed explanation of these three approaches is explained below.

3.3.1 VGG16

VGG extracts spatial information effectively from facial images and has various convolutional and fully connected (FC) layers that capture meaningful patterns that are informative to recognize facial expressions. VGG16 has 10 identical deep neural structures which are determined from Multi-Layer Perceptron (MLP) network and VGG16 model. The VGG16 and every 10 identical networks of deep neural with convolutional block, i.e., preserve bias and weight of removing FC layer and VGG16 model. MLP input structure of neural network outputs to build a classification network. More images from different classes are correctly recognized by utilizing VGG16. It has 16 layers with convolutional and max-pooling layers followed by FC layers. VGG16 has 4096 features and efficiently captures complex facial features by utilizing its pre-trained weights on an image. The features are extracted from the FC7 layer which has spatial information. Its hierarchical structure enables it to learn hierarchical representations of facial expressions which helps in accurate recognition of facial images.

3.3.2 ResNet 18

ResNet18 effectively extracts both low-level and high-level features in facial images via pooling layers. ResNet has 512 features and makes efficient deep network training which allows it to capture variations and complex patterns in facial expressions. The initial layers capture the features of low-level like texture, edges, and their shapes. Residual connections make the preservation of data from prior layers which allows for an extraction of high-level features like emotions, landmarks, and facial expressions. It contains 5 convolutional structures, FC layers, and a SoftMax activation layer. The initial convolutional structure has a 1D convolutional layer, activation layer, and batch normalization. The features are extracted from pooling 5 layers which have low dimensions compared to FC because of global average pooling layer that assists in minimizing the feature representations of dimensionality. The amount of convolution kernels in a 1D convolutional layer is 64, convolution kernel size and padding modes are 14, and Rectified Linear Unit (ReLU) is utilized as an activation layer. The 2^{nd} to 5^{th} structures of convolution are the same form which has a feature map and convolutional block but the amount of convolutional kernels varies for blocks. It enables efficient extraction of hierarchical features with low and high-level features which increases the accuracy of FER.

3.3.3 AlexNet

It is a kind of CNN architecture that demonstrates the DL technique's effectiveness in image recognition tasks. It extracts local and global features in facial images and these features are extracted from the FC7 layer which captures discriminative facial features like eyebrows, eyes, mouth, and nose. The relative positions and spatial arrangements of these landmarks contain significant data about facial expressions that are learned via convolutional layers. Alexnet has 512 features and has 8 learning layers in that 5 are convolutional, 3 are maximum pooling layers, and 3 are FC layers with pooling and convolutional layers. It employs ReLU activation function instead of Tanh and Sigmoid activation function for all FC and convolutional layers. However, hence the value domain acquired after the ReLU has no interval. Therefore, Local Response Normalization (LRN) is employed that normalize data acquired from ReLU to conceal small neurons and enhance the ability of generalization in the model. By using these three techniques of VGG16, ResNet18, and AlexNet, various types of features like texture, patterns, and spatial arrangements are captured which maximizes the recognition accuracy in facial expressions. Overall, VGG16, ResNet18, and AlexNet have 8704 features. After extracting the features, the feature selection process is employed to select the most appropriate features in facial expressions.

3.4 Feature Selection

The ICO technique is utilized to choose the most relevant features that is 6552 from extracted features in FER which assists in minimizing the data dimensionality that enhances model efficiency and generalization. Recognizing the most information features increases the ability of a model to discriminate among various facial expressions. A brief explanation of this approach is discussed below.

3.4.1 Coati Optimization Algorithm (COA)

COA is a metaheuristic optimization approach inspired by coati's cooperative hunting behavior by employing cooperation between individuals to increase exploitation and exploration in search space which helps in addressing optimization issues effectively. COA can efficiently balance exploitation and exploration, adapt to dynamic environments, and effectively search optimal solutions with search space of high-dimensional images.

3.4.1.1 Initial Population

In COA, every coati has an independent location in search space and this location is among lower and upper boundary which is expressed in equation (2)

$$Pos = \begin{bmatrix} Pos_1 \\ \vdots \\ Pos_i \\ \vdots \\ Pos_N \end{bmatrix}_{N \times dim} \tag{2}$$

Where Pos indicates coatis position, N represents number of coatis, and m determines dimensions. In every analysis within COA, fitness of present position is stored. A fitness value of present location is displaced with enhanced fitness value whether fitness of following location is better than present location. However, whether next analysis cannot surpass the present location fitness, the present location's fitness value is maintained. This technique establishes best fitness

value so far is maintained via an optimization procedure. The COA goal is to converge towards a best possible solution by updating value of fitness while better solutions are found. This method of updating and storing fitness values makes COA track processes and maintain best fitness value acquired in optimization process.

3.4.1.2 Coati prey search and aggressive behavior (Exploration phase)

In initial phase, coatis are generated randomly in search space where half of a coatis scramble a tree to frighten iguanas. Then, half of them are below a tree awaiting for an iguana to fall on ground and attack it. Based on equation (3) updating coati's position climbing the tree. Iguana falls to a random location in search space while coatis scare iguana. Their location is updated based on a comparison among fitness of the present coati position and optimal fitness which is expressed in equation (4)

$$Prey: Prey_j = lb_j + rand\ (0,1).\left(ub_j - lb_j\right), j = 1, 2, \dots, m \quad (3)$$

$$Pos_{i,j}^{p1}: Pos_{i,j}^{p1} = \begin{cases} Pos_{i,j} + rand\ (0,1).\left(Prey_j - I.Pos_{i,j}\right) & F_{prey} < F_i \\ Pos_{i,j} + rand\ (0,1).\left(Pos_{i,j} - Prey_j\right) & else \end{cases} for\ i = \frac{N}{2} + 1, \frac{N}{2} + 2, and\ j = 1, 2, \dots, m \quad (4)$$

Where $Prey$ indicates coatis location in tree, j determines dimension, P stands for initial stage, Pos_i illustrates present location, and i represents random integer generated in a set [1, 2]. F_{prey} indicates computing the fitness of present position. Whether the present fitness is less than optimal fitness, it indicates present coati location is better and its position is updated. A relevant value of present location is computed which is expressed in equation (5)

$$Pos_i = \begin{cases} Pos_i^{p1}, F_i^{p1} < F_i \\ Pos_i, else \end{cases} \quad (5)$$

By using the above equation, the present position is calculated effectively in search space of coati.

3.4.1.3 Coati Escaping Behavior of Predator (Exploitation Phase)

The next stage is formulated by determining coati's natural fleeing behavior from predators. Based on equations (6) and (7), the present coati location is updated. While the coatis are attacked by predators, they flee from their present position. Hence, coatis utilize predator avoidance strategy to maintain them in a nearby safe location to their present position. This approach determines a ability of Coati to establish local searches.

$$Ib_j^L = \frac{lb_j}{t}, ub_j^L = \frac{ub_j}{t}, t = 1,2, \dots, T \quad (6)$$

$$Pos_i^{p2}: Pos_{i,j}^{p2} = Pos_{i,j} + \left(1 - 2 \times rand(0,1)\right).\left(Ib_j^L + rand(0,1).\left(ub_j^L - Ib_j^L\right)\right)\ i = 1, 2, \dots, N\ and\ j = 1, 2, \dots, m \quad (7)$$

Where $p2$ indicates next stage, Ib_j^L and ub_j^L represents local upper and lower bounds in j dimension. lb_j and ub_j determines upper and lower bounds of j dimension. T illustrates maximum amount of evaluations, Pos_i^{p2} indicates a coati new location that is computed in

second phase, and $Pos_{i,j}^{p2}$ refers to j dimension's location. Equation (8) computes a present fitness position. Whether a present fitness is lower than optimal fitness, the coati present position is better else, they remain unchanged in present position.

$$Pos_i = \begin{cases} Pos_i^{p2}, F_i^{p2} < F_i \\ Pos_i, else \end{cases} \quad (8)$$

Where F_i^{p2} determines computed present position fitness, Pos_i^{p2} represents present position, and Pos_i refers updated position. However, the COA has disadvantages like convergence speed, easily falling into local optima, and minimizing the ability in global exploration. To solve this issue, ICOA is utilized in this research.

3.4.2 Improved Coati Optimization Algorithm (ICOA)

The conventional COA is improved by using three approaches. They are search and encirclement approach by transmitting sound achieves convergence speed. Physical exertion-based escape approach is used for escaping while being chased which helps to not to fall from local optima. At last, lens-opposition learning approach is employed to explore greater range that increase global exploration capability. A detailed explanation of this approaches are discussed below.

3.4.2.1 Search and Encirclement Approach by Transmitting Sound

3.4.2.1.1 Search Approach

In COA, coati's approach their prey gradually which minimizes search range and their speed of movement slows down. This leads to decrease in convergence speed. Hence, search and encirclement model is established by transmitting sound for more rapidly locating their prey which increases convergence speed. ICOA employs acoustic signals for communication while coatis are engaged in cooperative hunting. By determining a signal strength emitted by prey, coati analyze the prey location and approach it. This procedure is called sound transmission search approach. The coati in searching for prey in trees depends on sounds generated by iguanas. Equation (9) calculates the time it considers for traveling signals from iguana to coati. Equation (10) estimates received intensity of sound on both right and left sides. At last, new coati location is acquired by using equation (11)

$$D = |\frac{Pos_{i,j} - Prey_j}{v}| \quad (9)$$

$$Di = \cos(2 \times \pi \times D \times P) \times a \quad (10)$$

$$Pos_i^{p1} = Pos_{i,j} + rand(0,1).\left(Pos_{i,j} - Pos_j\right).a.C.Di \quad (11)$$

Where $p1$ indicates initial phase, D represents time taking for traveling sound, v denotes sound transmission speed, $Pos_{i,j} - Prey_j$ refers to distance among iguana and coati, Di illustrates sound intensity. A parameter P determines a linearly minimizing number in [1, 3], a influence factor in transmitting a signal that minimizes amount of evaluations enhances, and C denotes control coefficient. These equations and parameters are utilized in ICOA is to compute coati positions with received acoustic signals in searching process.

3.4.2.1.2 Encirclement Approach

Once iguana is descend and scared to ground, coati in tree transmits signal sound to others on ground. They obtains signal and space search in sound source direction from numerous directions. Equations (12) to (14) are included in COA as a condition to update the location while coati surrounding a prey. A search direction is evaluated based on (12). This equation helps coati to move forward source sound. Moreover, integrated with equation (12), the new coati position is updated by utilizing equations (15) and (16). It provides coati location adjustment based on data acquired from sound signals.

$$DI = \cos\left(2 \times \pi \times D \times f\right) \tag{12}$$

$$\omega = 1 - \frac{e^{\frac{t}{T}-1}}{e-1} \tag{13}$$

$$JL = Pos_{i-1,j} - Pos_i^{p1}, jl = JL.R \tag{14}$$

$$Pos_i^{p1} = Pos_{i-1,j} + rand\ (0,1).(Pos_j - I.Pos_{i-1,j}) \tag{15}$$

$$Pos_i^{p1}: Pos_{i,j}^{p1} = \begin{cases} Pos_{i-1,j} + rand(0,1).(Pos_{i-1,j} - Prey_j).a.C.DI, jl \\ < JLPos_{i-1,j} + rand(0,1).(Pos_{i-1,j} - Prey_j), else \end{cases} \tag{16}$$

Where DI indicates prey's position, f determines number that minimizes linearly from 2 to 1. ω denotes adaptive weight, T indicates overall evaluation exponential, and t illustrates no. of evaluations. JL refers distance among two coatis, Jl indicates distance among 2 coati after sound signal is received, R determines random number among [-1, 1, V represents parameter, Pos_i^{p1} determines updated location, and $Pos_{i-1,j}$ indicates prior coatis location.

3.4.2.2 Physical Exertion Approach

Coati faces challenges in quickly finding secure positions nearby during pursuit of natural predators. Hence, physical exertion-based escape approach is constructed which enables the coati to have a narrow escape range choice while being chased. It increases ability to not to fall on local optima. During ICOA's exploitation stage, the coati is determined by fleeing under predators' pursuit. The coati run fastly and consider refuge in caves, jungle, and other terrain. While a predator's physical strength is gradually less than of coati's, the pursuits are abandoned. Hence, equation (18) is used to compute updated location. Else, a new location is produced based on equation (17)

$$Pos_i^{p2} = Pos_{i,j} + (1 - 2 \times rand\ (0,1).(Ib_j^L + rand(0,1).(ub_j^L - Ib_j^L) \tag{17}$$

$$Pos_i^{p2} = Pos_{i,j}.(2 \times rand\ (0,1) - 1).|E_1 - E|.\varepsilon \tag{18}$$

Where E_1 and E denotes coati and predator's physical strength parameters and their value of two parameter is randomly generated among [0, 1], and ε represents linearly minimizing number from 1 to 0 respectively.

3.4.2.3 Lens Opposition-based Learning Strategy

It generates present position in reverse order that enables it to explore greater range which increase the capability of global exploration. This approach integrates the reverse learning principle and lens imaging. By estimating present location coordinates and producing a reverse solution, searching scope is increased, and best locations are determined in search procedure. A convex lens is utilized for evaluating lens image; an image on another side is A*. It is evaluated on x-axis to x* and h* height. The backpropagation x* of x is acquired which is expressed in equation (19)

$$\frac{\frac{ub+lb}{2}-X}{X*-(ub+lb)/2}=\frac{h}{h*} \tag{19}$$

An lens opposition-based learning is acquired which is formulated in equation (20)

$$X_j^*=\frac{ub_j+lb_j}{2}+\frac{ub_j+lb_j}{2k}-\frac{X_j}{k} \tag{20}$$

Where k denotes location of an individual in dimension j and X_j^* indicates inverse solution of X_j, ub_j and lb_j determines maximum and minimum bounds of j dimension respectively. In conclusion, each of the three approaches contributes distinctively by improving the efficiency of COA. A search and encirclement approach is performed by transmitting sound which achieves convergence speed and enhances FER effectiveness. A physical exertion-based escape approach is used for escaping while being chased which helps to not fall from local optima and increase the model's adaptability in facial recognition tasks. A lens-opposition learning approach explores a greater range of facial expressions that increases global exploration capability and makes more understanding of different emotional states. Then, the selected features are fed as input into classification to recognize facial emotion expressions.

3.5 Classification

After selecting features, the ELU-GRU with layer normalization is used in this stage to recognize facial emotion expressions. The input data is in a temporal manner and it has spatial information, a number of frames are utilized for FER. GRU has the capability to capture both spatial and temporal dependencies effectively. Therefore, GRU classifier is used to recognize facial expressions. Even Recurrent Neural Unit (RNN) and Long Short-Term Memory (LSTM) can do the same. However, RNN has memory issue and LSTM has complex structure and large hyperparameters compared to GRU. By utilizing GRU, the space complexity is effectively reduced. ELU is an activation function that assist the model to learn robust representations. ELU helps to solve the vanishing gradient issue which allows the model to learn complex representations. Layer normalization is used to normalize the layer parameters which solves exploding gradient issue during training. It helps to maintain stable distribution of activation function throughout the network.

GRU utilize a gating mechanism of RNN to solve integral deficits and increasing the ability in learning dependencies in long-period of conventional RNNs. GRU has two gates that is reset gate r_t and update gate z_t. An x_t indicates input at time t, h_{t-1} represents hidden state at $t-1$ time. Furthermore, W_r and U_r determines weight matrix for input and hidden state. The forget and input gates are constructed with an update and reset gates z_t and r_t is employed to h_{t-1} to obtain h_t (candidate state). The mathematical formula for GRU gates are expressed in equations (21) to (24).

$$z_t = \sigma(W_t x_t + U_z h_{t-1} + b_z) \quad (21)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (22)$$

$$\widetilde{h_t} = \tanh(W\widetilde{hx_t} + U\tilde{h}(r_1 \times h_{t-1}) + b\,\tilde{h}) \quad (23)$$

$$h_t = (1 - z_t) \times h_{t-1} + z_t \times \widetilde{h_t} \quad (24)$$

Where σ and tanh determines logistic sigmoid and function of hyperbolic tangent, $\times$ indicates element-wise multiplication, and b is a bias vector. Then, the output from FC layers to output layer is formulation in equation (25)

$$y_t = ELU\ (W_\circ h_t + b_\circ) \quad (25)$$

Where y_t indicates output vector and ELU represents Exponential Linear Unit. The σ and tanh functions are expressed in equations (26)

$$\sigma(x) = 1/(1 + e^{-x}) \quad (26)$$

$$\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x}) \quad (27)$$

The ELU is formulated in equations (28) and (29)

$$f(x) = \begin{matrix} x & (x > 0) \\ \propto (e^x - 1) & otherwise \end{matrix} \quad (28)$$

Where $\propto$ defines hyper-parameter. ELU enables the mean activations which is near to 0, and speed by learning and ELU acquire greater accuracy than ReLU.

$$f(x) = \begin{matrix} x & (x > 0) \\ 0 & otherwise \end{matrix} \quad (29)$$

Finally, layer normalization is utilized to normalize the layer parameter and it effectively solves gradient exploding issue which maintains stable distribution in activation via networks. ELU-GRU with layer normalization not only captures temporal dependencies but also provides stable distribution. Through this technique, the recognition accuracy is effectively enhanced in facial emotion expressions.

4. Results

The proposed ELU-GRU with layer normalization approach is simulated by utilizing MATLAB (R2020b) with 16 GB RAM, INTEL i5 processor, Windows 10 operating system, 6GB GPU, and 1TB HDD. The proposed approach utilizes performance metrics like accuracy, sensitivity, specificity, f1-score, and Matthews Correlation Coefficient (MCC). These metrics mathematical formula are expressed in equations (30) to (34)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (30)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (31)$$

$$Specificity = \frac{TN}{TN+FP} \quad (32)$$

$$F1 - Score = \frac{2\times TP}{2\times TP+FP+FN} \times 100 \quad (33)$$

$$MCC = \frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (34)$$

Where TP represents True Positive, FP indicates False Positive, FN denotes False Negative, and FP refers False Positive respectively.

4.1 Performance Analysis

The performance analysis of ELU-GRU with layer normalization is represented in Tables 1 to 8. Table 1 and 2 denotes different optimization techniques utilizing AffectNet, CK+ datasets. The performance of Artificial Bee Colony (ABC), FruitFly Optimization Algorithm (FOA), Grey Wolf Optimization (GWO), and COA are compared with ICOA technique. The ICOA achieves higher performance compared to ABC, FOA, GWO, and COA due to refined exploitation and exploration approach which makes better search space and effective convergence speed. Hence, it achieves high accuracy of 77.67% and 97.36% using both datasets compared to existing techniques.

Table 1. Different optimization technique utilizing AffectNet

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	MCC (%)
ABC	74.73	74.69	72.31	73.65	70.95
FOA	72.08	71.51	70.55	72.34	72.76
GWO	67.55	65.64	63.86	65.82	74.46
COA	75.02	73.35	72.87	73.35	74.95
ICOA	**77.67**	**75.88**	**73.43**	**76.43**	**76.97**

Table 2. Different optimization technique using CK+

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	MCC (%)
ABC	96.12	95.56	94.79	93.86	90.54
FOA	90.24	89.91	88.10	89.77	88.44
GWO	95.13	93.91	94.36	93.35	92.42
COA	96.45	95.20	95.47	94.33	93.64
ICOA	**97.36**	**97.38**	**98.12**	**96.76**	**96.61**

Table 3 and 4 shows performance of K-fold analysis using AffectNet, CK+ datasets. K=5-fold cross-validation exhibits greater performance of 77.67% and 97.36% for both datasets due to its balanced trade-off among bias and variance. With 5 folds, the data is split into smaller which has more representative subsets for validation and training. This enables efficient performance by minimizing variability. Also, it assists in generalization performance to unseen data by training and validating iterative on various subsets.

Table 2. Performance of K-fold analysis using AffectNet

K-fold	Accuracy	Sensitivity	Specificity	F1-score	MCC

	(%)	(%)	(%)	(%)	(%)
3.00	73.13	74.8	71.17	72.49	74.92
5.00	**77.67**	**75.88**	**73.12**	**76.43**	**76.97**
7.00	73.18	73.97	70.12	74.30	74.87
9.00	72.56	71.54	70.97	72.32	73.98

Table 4. K-fold evaluation utilizi ELU-GRU with layer normalization ng CK+

K-fold	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	MCC (%)
3.00	95.61	94.85	96.04	94.88	95.75
5.00	**97.36**	**97.38**	**98.12**	**96.76**	**96.61**
7.00	95.35	94.77	96.06	95.55	94.46
9.00	94.37	94.95	95.47	95.47	93.68

Table 5 and 6 represents analysis of actual features using AffectNet, CK+ datasets. The performance of RNN, LSTM, Stacked_GRU, and Bi-GRU are compared with ELU-GRU with layer normalization. When compared to these existing techniques, ELU-GRU with layer normalization achieves high performance due to it has capability to reduce vanishing gradient issue in RNN and LSTM which results in enhanced long-term dependency. Also, ELU and layer normalization generates smoother gradients which maintains stable distribution. Therefore, ELU-GRU with layer normalization achieves high accuracy of 70.83% and 90.23% using both datasets respectively.

Table 5. Analysis of actual features using AffectNet

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	MCC (%)
RNN	57.89	56.16	55.89	57.51	56.65
LSTM	48.99	48.77	48.99	47.97	48.43
Stacked_GRU	43.21	37.31	42.71	42.58	40.23
Bi-GRU	69.09	70.70	67.24	67.62	67.45
ELU-GRU with layer normalization	**70.83**	**71.67**	**70.39**	**72.31**	**72.72**

Table 6. Analysis of actual features using AffectNet

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	MCC (%)
RNN	84.75	83.53	85.53	83.98	84.85
LSTM	87.92	86.54	85.32	86.42	85.70
Stacked_GRU	86.75	86.86	85.86	85.87	84.33
Bi-GRU	89.69	83.47	80.87	85.15	83.05
ELU-GRU with layer normalization	**90.23**	**91.05**	**92.19**	**90.60**	**92.09**

Table 7 and 8 determine analysis of optimized features using AffectNet, CK+ datasets. The performance of RNN, LSTM, Stacked_GRU, and Bi-GRU are compared with ELU-GRU with layer normalization. Figure 4 and 5 shows a graphical representation of optimized features analysis using AffectNet, CK+. The proposed technique achieves better accuracy due to it solves the exploding gradient issue during training captures both spatial and temporal data, and normalizes layer parameters which maintains stable distribution compared to existing techniques. Hence, ELU-GRU with layer normalization achieves high accuracy of 77.67% and 97.36% using both datasets compared to RNN, LSTM, Stacked_GRU, and Bi-GRU.

Table 7. Analysis of optimized features using AffectNet

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	MCC (%)
RNN	62.88	60.66	61.33	62.42	63.54
LSTM	52.36	50.65	50.35	52.45	51.45
Stacked_GRU	45.75	44.63	43.78	45.43	45.09
Bi-GRU	73.52	72.58	70.87	71.15	70.32
ELU-GRU with layer normalization	**77.67**	**75.88**	**73.12**	**76.43**	**76.97**

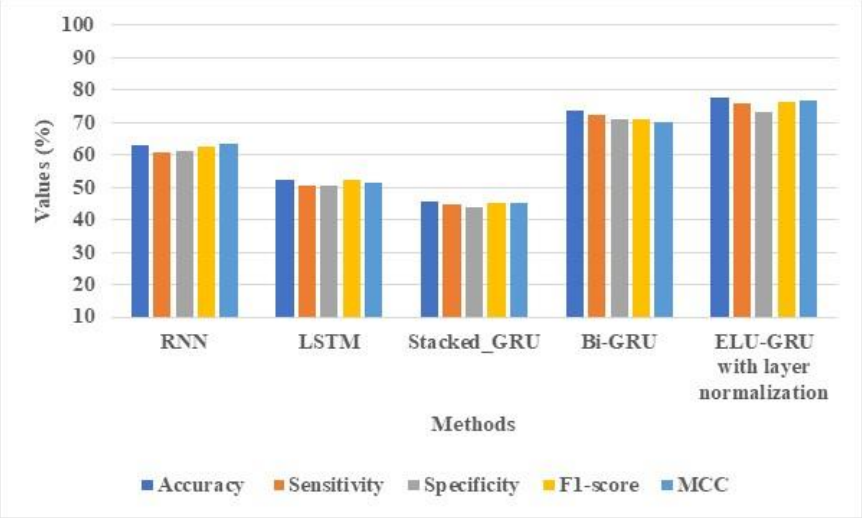

Figure 4. Graphical representation of optimized features using AffectNet

Table 8. Analysis of optimized features by employing CK+

Methods	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	MCC (%)
RNN	94.87	93.86	95.98	94.53	94.22
LSTM	96.76	95.54	96.32	94.75	93.49
Stacked_GRU	94.17	93.86	90.85	92.68	90.73

Bi-GRU	95.37	93.30	95.12	93.40	94.84
ELU-GRU with layer normalization	**97.36**	**97.38**	**98.12**	**96.76**	**96.61**

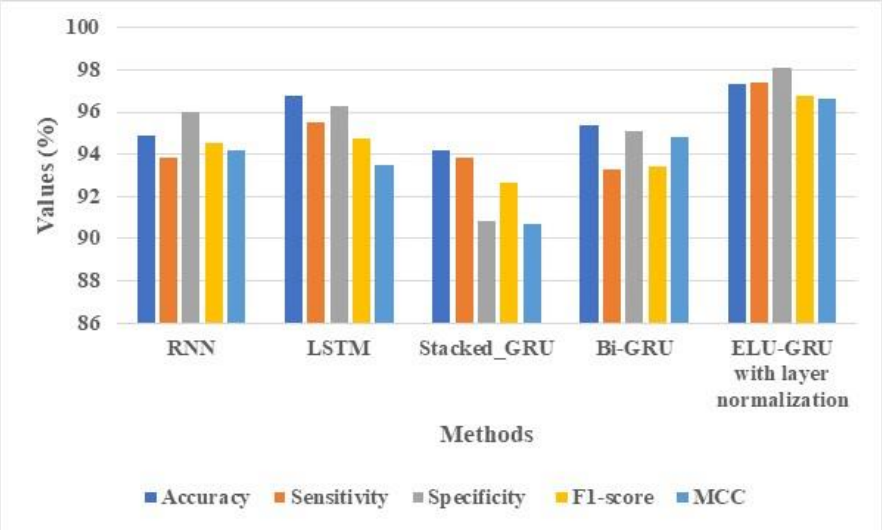

Figure 5. Graphical representation of optimized features using CK+

4.2 Comparative Analysis

Table 9 and 10 shows comparative analysis with existing techniques using AffectNet and CK+ datasets. The existing techniques like AMP-NET [16] and FER-DenseNet [17] are compared with proposed ELU-GRU with layer normalization using AffectNet. The GA-ELM [18], Multi-Attention [19], and CLBP-CNN [20] are the existing techniques compared with proposed approach using CK+ dataset. Due to proposed approach has capability to reduce vanishing gradient issue which results in enhanced long-term dependency. Also, ELU and layer normalization generates smoother gradients which maintains stable distribution. Hence, ELU-GRU with layer normalization achieves high accuracy of 77.67% and 97.36% for both datasets When compared to these five existing techniques.

Table 9. Comparative analysis with existing techniques using AffectNet

Methods	**Accuracy (%)**
AMP-NET [16]	61.74
FER-DenseNet [17]	59.38
Proposed ELU-GRU with layer normalization	77.67

Table 9. Comparative analysis with existing techniques using CK+

Methods	**Accuracy (%)**
GA-ELM [18]	93.53

Multi-Attention [19]	89.52
CLBP-CNN [20]	91
Proposed ELU-GRU with layer normalization	97.36

4.3 Discussion

In this section, benefits of ELU-GRU with layer normalization and disadvantages of existing techniques are discussed. The limitation of existing techniques are AMP-Net [16] suffers from over fitting issues because of limited training data which hinders the capability to generalize well on unseen facial expressions. FER-DenseNet [17] has low recognition accuracy in certain facial expressions like disgust and fear due to subtle and nuanced facial features which leads to challenges in detecting and classifying accurately. GA-ELM [18] lacks suboptimal performance because of GA's exploration limitations and ELM's potential challenges in capturing intricate relationships with multimodal data. The ICOA- ELU-GRU with layer normalization overcome these existing limitations. GRU captures both spatial and temporal data effectively. ELU helps to solve the vanishing gradient issue and layer normalization normalizes the layer parameter which maintains a stable distribution. By performing these processes, the proposed approach achieves accurate recognition in FER. Therefore, when compared to existing techniques like AMP-Net, FER-DenseNet, GA-ELM, Multi-Attention, and CLBP-CNN, the proposed approach achieves high accuracy of 77.67% and 97.36% for both datasets respectively.

5. Conclusion

The ICOA-ELU-GRU with layer normalization is proposed to recognize facial expressions. ICOA achieves convergence speed, avoid local optima, and global exploration ability with a help of search and encirclement approach, physical exertion-based escape model and lens opposition model which enhances the effectiveness in FER, makes more understanding of different emotional states. ELU-GRU with layer normalization effectively solves vanishing gradient issues, captures both spatial and temporal data, and normalizes layer parameters which maintain stable distribution in FER and recognize emotions accurately. By performing these operations, ELU-GRU with layer normalization achieves better accuracy of 77.67% and 97.36% for both datasets compared to AMP-Net, FER-DenseNet, GA-ELM, Multi-Attention, and CLBP-CNN.In the future, Machine Learning (ML) will be considered to further increase the model performances.

Funding Information: This research was no support from any funding agencies.

Conflict of Interest: The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

References

[1] Sharafi, M., Yazdchi, M., Rasti, R. and Nasimi, F., 2022. A novel spatio-temporal convolutional neural framework for multimodal emotion recognition. *Biomedical Signal Processing and Control*, *78*, p.103970.

[2] Middya, A.I., Nag, B. and Roy, S., 2022. Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. *Knowledge-Based Systems*, *244*, p.108580.

[3] Bian, Y., Küster, D., Liu, H. and Krumhuber, E.G., 2023. Understanding naturalistic facial expressions with deep learning and multimodal large language models. *Sensors*, *24*(1), p.126.

[4] Ni, R., Yang, B., Zhou, X., Cangelosi, A. and Liu, X., 2022. Facial expression recognition through cross-modality attention fusion. *IEEE Transactions on Cognitive and Developmental Systems*, *15*(1), pp.175-185.

[5] Zou, S., Huang, X., Shen, X. and Liu, H., 2022. Improving multimodal fusion with Main Modal Transformer for emotion recognition in conversation. *Knowledge-Based Systems*, *258*, p.109978.

[6] Wang, S., Qu, J., Zhang, Y. and Zhang, Y., 2023. Multimodal emotion recognition from EEG signals and facial expressions. *IEEE Access*, *11*, pp.33061-33068.

[7] Heredia, J., Lopes-Silva, E., Cardinale, Y., Diaz-Amado, J., Dongo, I., Graterol, W. and Aguilera, A., 2022. Adaptive multimodal emotion detection architecture for social robots. *Ieee Access*, *10*, pp.20727-20744.

[8] Chen, L., Wang, K., Li, M., Wu, M., Pedrycz, W. and Hirota, K., 2022. K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition in human–robot interaction. *IEEE Transactions on Industrial Electronics*, *70*(1), pp.1016-1024.

[9] Singh, R., Saurav, S., Kumar, T., Saini, R., Vohra, A. and Singh, S., 2023. Facial expression recognition in videos using hybrid CNN & ConvLSTM. *International Journal of Information Technology*, *15*(4), pp.1819-1830.

[10] Jabbooree, A.I., Khanli, L.M., Salehpour, P. and Pourbahrami, S., 2023. A novel facial expression recognition algorithm using geometry β–skeleton in fusion based on deep CNN. *Image and Vision Computing*, *134*, p.104677.

[11] Liu, S., Huang, S., Fu, W. and Lin, J.C.W., 2024. A descriptive human visual cognitive strategy using graph neural network for facial expression recognition. *International Journal of Machine Learning and Cybernetics*, *15*(1), pp.19-35.

[12] Filali, H., Riffi, J., Boulealam, C., Mahraz, M.A. and Tairi, H., 2022. Multimodal emotional classification based on meaningful learning. *Big Data and Cognitive Computing*, *6*(3), p.95.

[13] Yu, W. and Xu, H., 2022. Co-attentive multi-task convolutional neural network for facial expression recognition. *Pattern Recognition*, *123*, p.108401.

[14] Garcia-Garcia, J.M., Lozano, M.D., Penichet, V.M. and Law, E.L.C., 2023. Building a three-level multimodal emotion recognition framework. *Multimedia Tools and Applications*, *82*(1), pp.239-269.

[15] Boughida, A., Kouahla, M.N. and Lafifi, Y., 2022. A novel approach for facial expression recognition based on Gabor filters and genetic algorithm. *Evolving Systems*, *13*(2), pp.331-345.

[16] Liu, H., Cai, H., Lin, Q., Li, X. and Xiao, H., 2022. Adaptive multilayer perceptual attention network for facial expression recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(9), pp.6253-6266.

[17] Xie, Y., Tian, W., Zhang, H. and Ma, T., 2023. Facial expression recognition through multi-level features extraction and fusion. *Soft Computing*, *27*(16), pp.11243-11258.

[18] Pan, B., Hirota, K., Jia, Z., Zhao, L., Jin, X. and Dai, Y., 2023. Multimodal emotion recognition based on feature selection and extreme learning machine in video clips. *Journal of Ambient Intelligence and Humanized Computing*, *14*(3), pp.1903-1917.

[19] Zhi, J., Song, T., Yu, K., Yuan, F., Wang, H., Hu, G. and Yang, H., 2022. Multi-attention module for dynamic facial emotion recognition. *Information*, *13*(5), p.207.

[20] Mukhopadhyay, M., Dey, A. and Kahali, S., 2023. A deep-learning-based facial expression recognition method using textural features. *Neural Computing and Applications*, *35*(9), pp.6499-6514.

[21] AffectNet dataset link: https://github.com/djordjebatic/AffectNet

[22] CK+ dataset link: https://github.com/WuJie1010/Facial-Expression-Recognition.Pytorch

[23] Pei, X., hong Zhao, Y., Chen, L., Guo, Q., Duan, Z., Pan, Y. and Hou, H., 2023. Robustness of machine learning to color, size change, normalization, and image enhancement on micrograph datasets with large sample differences. *Materials & Design*, *232*, p.112086.

[24] Sharma, S., Guleria, K., Tiwari, S. and Kumar, S., 2022. A deep learning based convolutional neural network model with VGG16 feature extractor for the detection of Alzheimer Disease using MRI scans. *Measurement: Sensors*, *24*, p.100506.

[25] Zhao, Y., Zhang, X., Feng, W. and Xu, J., 2022. Deep learning classification by ResNet-18 based on the real spectral dataset from multispectral remote sensing images. *Remote Sensing*, *14*(19), p.4883.