



Multi-class Facial Emotion Expression Identification Using DL-Based Feature Extraction with Classification Models

M. Anand¹ · S. Babu¹

Received: 6 December 2022 / Accepted: 4 January 2024
© The Author(s) 2024

Abstract

Facial expression detection from images and videos has recently gained attention due to the wide variety of applications it has found in the field of computer vision such as advanced driving assistance systems (ADAS), augmented and virtual reality (AR/VR), video retrieval, and security systems. Facial terms, body language, hand gestures, and eye contact have all been researched as a means of deciphering and understanding human emotions. Automated facial expression recognition (FER) is a significant visual recognition procedure because human emotions are a worldwide signal used in non-verbal communication. The six primary universal manifestations of emotion are characterized as happiness, sadness, anger, contempt, fear, and surprise. While the accuracy of deep learning (DL)-based approaches has improved significantly across many domains, automated FER remains a difficult undertaking, especially when it comes to real-world applications. In this research work, two publicly available datasets such as FER2013 and EMOTIC are considered for validation process. Initially, pre-processing includes histogram equalization, image normalization and face detection using Multi-task Cascaded Convolutional Network (MT-CNN) is used. Then, DL-based EfficientNetB0 is used to extract the features of pre-processed images for further process. Finally, the Weighted Kernel Extreme Learning Machine (WKELM) is used for classification of emotions, where the kernel parameters are optimized by Red Fox Optimizer (RFO). From the experimental analysis, the proposed model achieved 95.82% of accuracy, 95.81% of F1-score and 95% of recall for the testing data.

Keywords Deep learning · Multi-task cascaded convolution network · EfficientNetB0 · Weighted kernel extreme learning machine · Red fox optimizer · Facial expression recognition

Abbreviations

ADAS	Advanced driving assistance systems
AR/VR	Augmented and virtual reality
FER	Facial expression recognition
MT-CNN	Multi-task cascaded convolution network
WKELM	Weighted kernel extreme learning machine
RFO	Red fox optimizer
HOG	Histogram of gradients
LRFE	Low-Resolution Facial Expression dataset (LRFE)
FL	Face localization
AMP	Adaptive multilayer perceptual

MFL	Multi-label focal loss (MFL)
V	Valence
A	Arousal
D	Dominance
AMT	Amazon Mechanical Turk
SLFN	Single-hidden-Layer Feedforward Neural Network
ERM	Empirical risk minimization
SRM	Structural risk minimization

1 Introduction

The human life and work are developed by the application called information technology using Artificial Intelligence (AI) technique. Nowadays, a typical person is developed with AI in the form of monitor screens, security control systems, commanding the functions for social work by using voices and interface between computers and humans called Human Recognition Face [1]. A filtering mechanism for

✉ M. Anand
am4639@srmist.edu.in
S. Babu
babus@srmist.edu.in

¹ Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur 603203, Tamil Nadu, India

EEG signals is used to shape information about human emotions, where face expression recognition system is one way to do this process [2]. Nonverbal communication, including facial expression, is a significant means of conveying one's emotions, intentions, objectives, and ideas to others [3, 4]. Facial appearance is the consequence of a facial sign or expression that reveals the location of the human face. Many computers today employ a method of estimating how satisfied a customer will be with a service based on the expressions portrayed on the customer's face. A customer's happiness level as shown by a selection made on a screen may not always be reliable [4]. System accuracy may be improved with the use of AI through real-time facial expression detection, and in that way, it can read customers' emotions in real time [5].

Six human emotions have been identified as universally perceivable without regard to cultural background by psychologists Ekman and Friesen [6]. Many research has used this fact to classify a set of six or seven emotions, with neutral expressions included as the seventh. Targets of study have broadened in recent years to include not just physical manifestations like sadness, pain, and fatigue, but also included mental manifestations like agreement, concentration, interest, thought, and confusion [7, 8]. The recognition of natural facial expressions, as opposed to the research through database including exaggerated reactions to the constrained context, is also the subject of active study. Despite these advances, FER technology is currently at a stage where it has limited usefulness [9].

There are three phases to the FER system used for expression recognition. Facial recognition is the initial process. This phase focuses towards locating a face in an input image and identifying its constituent parts. Adaboost, Haar-cascade and Histogram of Gradients (HOG) are all good examples of suitable algorithms [10]. The identified face is then subjected to either a geometric feature-based or an appearance feature-based extraction of its distinguishing characteristics. Emotions are then categorized using the method's retrieved characteristics in a classification phase [11]. The science of facial expression recognition relies heavily on data sets. Facial expression recognition is affected by two classes of variables. Distinct personal characteristics, such as gender, ethnicity, and age, constitute the first category. Lighting, postures, resolution, and noise are all examples of environmental influences that falls under second category. Although the first category of external influences had a greater impact in controlled circumstances, the external factors on second type were highly affected than first type [12]. The dataset has to be sufficiently rich to account for these considerations to solve this issue. For this reason, the study relied on data augmentation to fill in some gaps. Making use of many datasets simultaneously is another option. The goal is

to learn and validate hypotheses by merging datasets from the same domain.

To identify the facial emotions effectively, the research considered two different datasets such as FER2013 and EMOTIC to solve the issues which fall in second category. Three steps are provided in the research work. The validation analysis of proposed model with existing techniques are carried out by using different metrics. The main contribution of the work is described as follows:

1. Face detection process is approved out by using MTCNN model, where other pre-processing techniques such as normalization and histogram equalization are used to improve the input images that solves the issues of lighting, poor illuminations, etc.
2. Features are extracted from the pre-processed images using Deep learning-based model called EfficientNetB0, that is used to drop the unwanted features and minimized the computational time by changing the filter sizes and used only convoluted features of EfficientNet for efficient feature extraction.
3. Classification of emotions are carried out by using weighted KELM, where kernels are optimally selected by RFO.

The remaining paper is constructed as: Sect. 2 offers the study of existing works that are implemented on FER2013 and EMOTIC datasets. The brief explanation of proposed model with mathematical expression is given in Sect. 3. The validation analysis of proposed model with existing techniques in terms of different metrics is depicted in Sect. 4. Finally, the scientific contribution with future scope is designated in Sect. 5.

2 Related Works

Chowdary et al. [13] uses transfer learning techniques to address the issues of emotion recognition. Pretrained networks from the Resnet50, vgg19, Inception V3, and Mobile Net frameworks were used in this study. The study removed the pre-trained ConvNets' fully connected layers and replaced them with proposed layer, according to the task's instruction set size. The newly added layers were only trained to update their weights, which was the last restriction. The CK + database was used in the experiment, and on average, a 96% success rate was found for emotion recognition tasks.

A novel model, developed by Bodavarapu et al. [14], was able to effectively work with low-resolution, unreliable images. To conduct the research, the authors compiled a collection of FER images with varying degrees of resolution quality from a number of sources, creating the

low-resolution facial expression dataset (LRFE). A novel hybrid filtering approach was also presented, which combines the strengths of the Gaussian and Bilateral non-local filtering methods. Densenet-121 achieved an accuracy of 0.60 on FER2013 and 0.68 on LRFE. Densenet-121 and the hybrid filtering approach both led to an accuracy of 0.95. As a result of using the hybrid filtering technique, Resnet-50, MobileNet, and Xception models also showed promising results. When used in conjunction with the hybrid filtering technique, the suggested model improved accuracy to 0.85.

To improve accuracy, Sivaiah et al. [15] suggested Capsule Networks (CapsNet) for FER. However, the face images that were considered for training contain unnecessary information that slows down convergence and increases the number of iterations required to train a facial recognition system. Considering this, the study suggested integrating face localization (FL) with CapsNet in the model to filter out any irrelevant details in the facial images that might hinder the training process. The experiments were conducted on standard datasets including JAFFE, CK+, and FER2013; the results show that FL-CapsNet was superior to the currently-available CapsNet-based FER models.

The human visual system's facial perception process and its sensitivity to varied facial traits serve as inspiration for Liu et al. [16] to develop an Adaptive Multilayer Perceptual (AMP) Attention Network. Using several fine-grained features, AMP-Net was able to extract global, local, and salient facial emotional characteristics, allowing it to uncover the fundamental diversity and crucial information of facial expressions. AMP-Net's adaptive guidance of the network toward multiple finer and recognizable local patches that were resilient to occlusion and variable postures improves the network's capacity to learn probable facial variety information. The proposed global perception module learnt a wide range of receptive field features in AMP-Net boosts prominent face area characteristics with strong emotion correlation from past knowledge to improve texture detail capture and stop information loss. An extensive experiment showed that AMP-Net achieved state-of-the-art results and generalizability across a wide range of real-world datasets, such as RAF-DB.

To categorize FER, Chaudhari et al. [17] used the ResNet-18 model and transformers. In this research, the study compared the model's results on hybrid datasets with state-of-the-art models. This research details the fine-tuned transformer pipeline and accompanying techniques for face identification, cropping, and extraction using the latest deep learning models.

Hoang et al. [18] provides a method for predicting a person's emotional state based on the detection of visual associations between the nearby background objects. The study utilized both spatial and semantic qualities of objects in the scene to ascertain the relative importance of each contextual feature in determining the overall influence on

the focus of attention. The model incorporates the features with data on the environment and the subject's physical state to speculate on the subject's emotional state. The experiments were carried out on the CAER-S dataset with existing techniques to show the presentation of the projected model.

Fujisawa et al. [19] offer a model for emotion recognition based on combining the image features derived from emoticons with the text characteristics retrieved from tweets to form a feature vector input. Experiments confirming the suggested method's high accuracy reveal that the proposed model was superior than existing approaches that rely solely on text characteristics.

An alternative deep learning architecture for emotion identification was proposed by Bendjoudi et al. [20]. The architecture was composed of three primary components: a module for fusing the two sets of extracted features into a single decision. In addition, the synergy between loss functions was shown by comparing three categorical and functions for multi-task learning. Then, the work suggests a novel loss function based on the local loss to handle asymmetrical data, which was called as the multi-label focal loss (MFL). From the experiments on the EMOTIC dataset, the study concluded that MFL combined with the Huber loss exceeded the state-of-the-art on rare labels more effectively than any other combination. The proposed problem can also be closely correlated to detection approaches [31–33] where knowledge-based models have been found to be of high utility.

3 Proposed System

In this section, the three major steps are used, which is exposed in Fig. 1.

Figure 1 represents the proposed process for emotion prediction using a combination of datasets and deep learning techniques. Initially, two datasets, namely FER2013 and EMOTIC, are utilized. The data undergoes a pre-processing stage where MTCNN is employed for face detection, and then normalization and equalization are performed to standardize the images. Following this, feature extraction is carried out using the EfficientNetB0 model, which transforms the images into a set of features that are more suitable for the subsequent classification task. For the classification of emotions based on the extracted features, the Weighted Kernel Extreme Learning Machine (KELM) is used. Importantly, the selection of optimal kernels for this classification is determined by the Random Forest Optimizer (RFO). Once the model has been trained and the appropriate kernels are chosen, it is then capable of predicting emotions based on new input data.

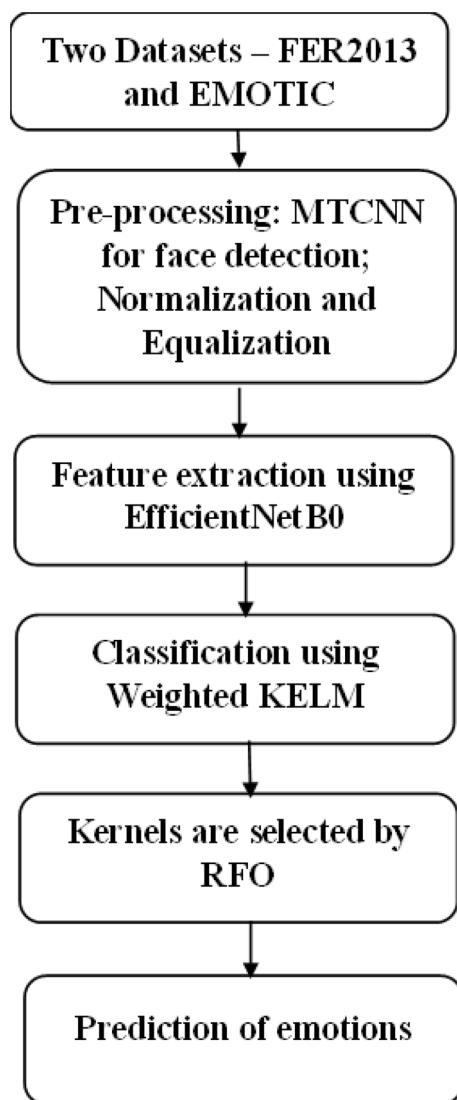


Fig. 1 Working flow of proposed model

3.1 Dataset Description

3.1.1 Facial Expression Recognition 2013 (FER-2013)

In 2013, Yang et al. presented the 2013 Facial Expression Recognition dataset (FER2013). This dataset is available through Kaggle. Figure 2 presents the sample images of this dataset.

Each face in the FER-2013 dataset has been assigned an emotion category, and the images are grayscale with a determination of 48 pixels by 48 pixels. Table 1 summarizes the index labels (0–6) used to categorize the 35,887 samples that make up the FER-2013 dataset. These labels are based on 7 (seven) distinct categories of micro expression.

3.1.2 Emotic Dataset

Images of individuals in natural settings, tagged with information on their apparent emotional states; this is the EMOTIC dataset. A total of 23,571 images and 34,320 persons annotations are included in the collection. The Google search engine was used to acquire some of the images. The researcher did this by using a set of queries that included terms for a wide range of locations, social contexts, activities, and mental states. All other images are from COCO [22] and Ade20k [23]—two publicly available benchmark datasets. The images depict a wide range of scenarios, with individuals in a variety of locations, groups, and activities. Three annotated photos from the EMOTIC dataset are displayed in Fig. 3.

Amazon Mechanical Turk (AMT) was used for the image annotation. Annotators were tasked with assigning labels the images based on their interpretations of the subjects' emotions. Figure 3b depicts a child who would interest in eating a chocolate than an apple. As a result of his intentions are considered as one of uncomfortable feeling and irritation. The continuous dimensions Valence (V), Arousal (A), and Dominance (D) as shown in Fig. 3 are also used to comment the images. In Fig. 3c, for instance, the individual is engaged in a task that implies anticipation that appears to be enjoying himself and is completely absorbed in what he is doing at this moment.

After the initial round of annotations, the study split the dataset into three groups, with each set including a comparable distribution of emotional categories across training (70%), validation (10%), and testing (30%) images. Additionally, four annotators are used for validation and two annotators are used for testing images, respectively. Some images are noisy and therefore, it is removed and then, three annotators are used for testing set.

3.2 Pre-processing

With multi-task learning and a convolutional neural network (CNN), MTCNN builds an outline for combining face detection and alignment. As can be seen in Fig. 4, the suggested CNNs comprise of three distinct stages: the P-net, the R-net, and the O-net.

The initial process begins by making a pyramid out of the input image by scaling it to various sizes. The images are then analyzed using P-net, which looks for b-boxes. After that, the study utilizes NMS to merge all the duplicated candidates. R-net, which further eliminates non-face windows, is fed the output of the first stage. In a similar way, potential NMS candidates are combined. Aside from exchanging the R-net for the more durable O-net, the third stage is virtually indistinguishable from the second. In addition to the finished b-box, five landmarks are created.



Fig. 2 Image in Dataset FER-2013

Table 1 Sum of data in the FER-2013 dataset

Micro-expression	Training dataset	Dataset total	Validation data Public	Validation data Public
Angry	3995	4953	467	467
Happy	7215	8989	895	895
Sadness	4830	6077	653	653
Surprise	3171	4002	415	415
Disgust	436	547	56	56
Fear	4097	5121	496	496
Contempt	4965	6198	607	607
	28,709	35,887	3589	3589

The MT-CNN in the proposed work consists of three stages:

P-net (proposal network): this is the first stage. The P-net proposes candidate facial regions from the image. It rapidly scans the image to identify potential face candidates by resizing and cropping various regions.

R-net (refine network): the R-net processes the candidate regions from the P-net to reject false positives and further refine the bounding boxes for potential faces.

O-net (output network): this is the final stage. The O-net considers the regions passed on from the R-net to finalize the bounding boxes and additionally perform facial landmarks localization, pinpointing features like eyes, nose, and mouth.

3.2.1 Histogram Equalization

Histograms for pixel-by-pixel in grey-level mapping have allowed for successful image improvement, but adaptive histogram equalization grounded on probability theory has shown to be superior in achieving improvement on input images.

3.2.2 Image Normalisation

Data normalization's primary goal is to convert a dataset's numeric columns to a more human-readable scale while maintaining the original data's range variability. Data normalization is optional in machine learning, but important

Fig. 3 Model images in the EMOTIC along with their comments

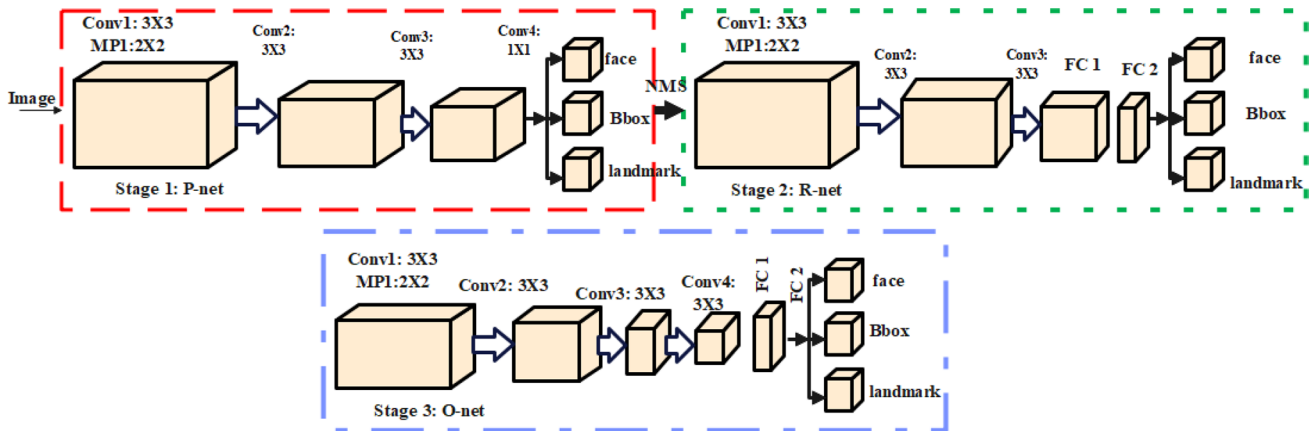
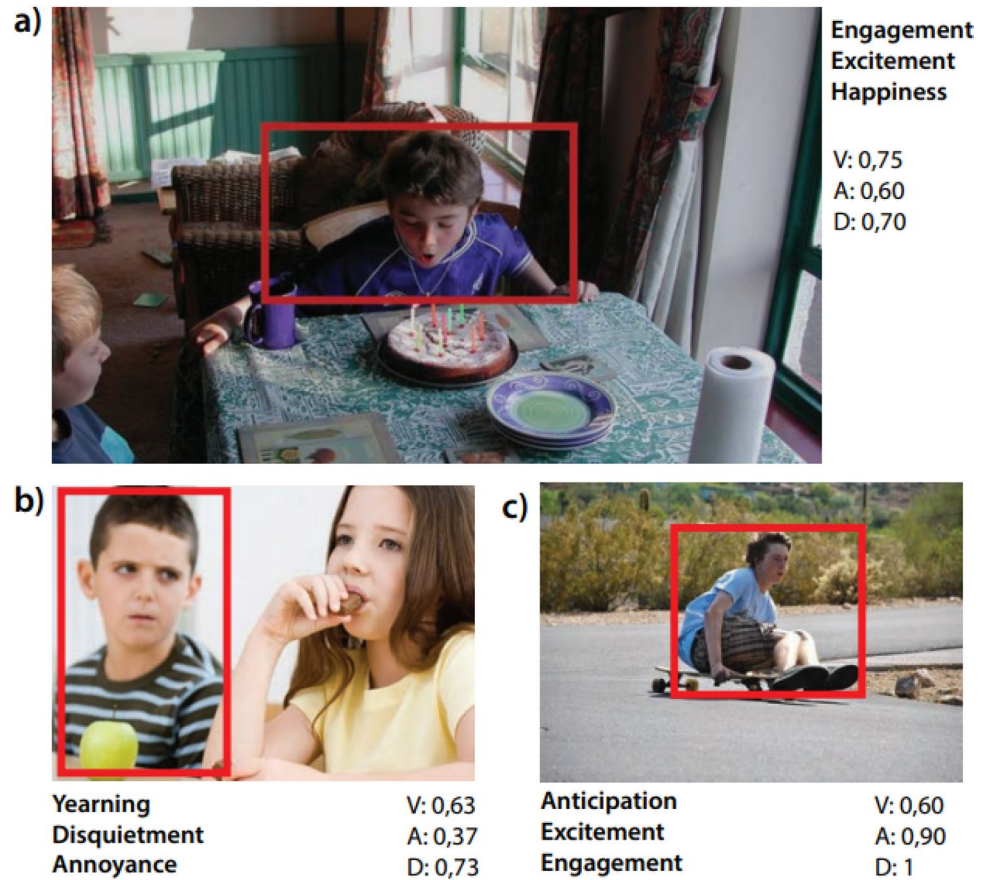


Fig. 4 Architecture of MT-CNN

when features have a wide range of values. By using data normalization, we can ensure that each row only stores the information it needs once and cut down on redundant records. Using data normalization, the image's irrelevant information is removed and noises are removed. Normalization of data is represented mathematically in Eq. (1).

$$EF_{st}^{\text{normalization}} = (ef1 - ef2) * \frac{(EF_{st} - EF^{\min})}{(EF^{\max} - EF^{\min})} + ef1. \quad (1)$$

The attribute of the data to be normalized is denoted by EF_{st} in Eq. (1), and the resulting normalized data are referred to as $EF_{st}^{\text{normalization}}$, the maximum characteristic



Fig. 5 Architecture of feature extraction model

value related to each record is signified as EF^{\max} , and EF^{\min} refers to the minimum attribute value related to each record.

3.3 EfficientNet Feature Extraction

Models employed in this dataset are increasingly complicated since 2012, although many of these models are inefficient in terms of the amount of processing power they need. The EfficientNet model may be part as a class of CNN models, placing it among the state-of-the-art replicas. While the number of estimated restrictions does not drastically rise with the number of replicas in the EfficientNet group (which contains of 8 replicas among B0 and B7), accuracy does improve.

Deep learning constructions are designed to discover more effective methods with less model complexity. EfficientNet outperforms competing state-of-the-art models because it scales down the model equally in depth, and resolution without sacrificing accuracy. In order to determine how various scaling dimensions of the baseline network relate to one another while working with a given set of resources, the first stage of the compound scaling approach is to look for a grid. Scaling factors for depth, width, and resolution may be calculated in this fashion. Next, the target network is scaled up from the original network using these coefficients. Figure 5 presents the architecture of this model [24].

Although the inverted bottleneck MBConv was originally introduced in MobileNetV2, EfficientNet makes significantly greater use of it than MobileNetV2. MBConv are composed of an expansion layer followed by a compression layer; this design allows for direct connections between bottlenecks, which link much fewer channels than extension layers. In comparison to conventional layers, this architecture's deep separable convolutions cut down on computation by almost a factor of k^2 , where k is the size of the kernel, denoting the width and height of the 2D convolution window.

The compound coefficient is applied according to the formula in Eq. (2) scale in a consistent manner across depth, breadth, and resolution.

$$\text{Depth : } d = a^{\psi}$$

$$\text{Width : } w = \beta^{\psi}$$

$$\text{Resolution : } r = \gamma^{\psi}$$

$$a \geq 1, \beta \geq 1, \gamma \geq 1,$$

(2)

where, $\alpha, \beta, \gamma, \alpha, \beta, \gamma$ constants that may be found with grid search. The number of resources available for model scaling is determined by the parameter, while $\alpha, \beta, \gamma, \alpha, \beta, \gamma$ find out how the network's depth, width, and resolution are affected by the additional resources. FLOPS are directly related to d, w^2, r^2 in a standard convolution procedure. When scaling a convolution network according to Eq. (2), the network's FLOPS rise by a factor of around $(\alpha, \beta^2, \gamma^2)^{\varphi} (\alpha, \beta^2, \gamma^2)^{\varphi}$. This is because the majority of the network's FLOPS are spent on convolution operations, which are expensive.

The compound scaling approach scales this model in two stages:

First, we assume that there are twice as numerous resources and do a grid search with $\varphi = 1$ to get the optimal values for α, β, γ .

Second, using Eq. (2), the method scale up the initial network to get EfficientNet-B1 through B7, fixing the values of α, β, γ .

3.4 Classification by Regularized/Kernel Extreme Learning Machine (RELM)

In this study, we explore the use of RELM [25] to identify the facial expressions from the extracted features. Here, the study uses a RELM classifier to predict the emotions that is currently displaying based on the above feature extraction model. Fast training times and strong generalization performance are two of ELM's strengths and form as a single-hidden-layer feed-forward neural network (SLFN). Nonetheless, it is founded on the notion of empirical risk minimisation (ERM), which

often leads to overfitting. RELM, which is founded on structural risk minimisation (SRM), was developed to address this shortcoming. Reduced training error on the training data but increased testing error on unknown data (overfitting) result from a combination of a minimal empirical risk and a limited data set. As a result, RELM operates according to the SRM principle, which in turn is founded upon the statistical learning theory. It offers the bound-on generalizability, which is the link between empirical risk and actual risk. In this case, the square root of the error, or $\|\epsilon\|^2$, represents the empirical risk, whereas a square root of the deviation, or, represents the structural risk as $\|\beta\|^2$.

Specifically, there are n distinct training samples $(x_i, t_i) \in \mathbb{R}^k * \mathbb{R}^m$. It uses $g(x)$ as its activation function. The RELM with N hidden nodes is modeled for the i th sample as

$$\sum_{i=1}^{\tilde{N}} \beta_i g_i(x_q) = \sum_{i=1}^{\tilde{N}} \beta_i g_i(w_i x_q + b_i), \quad (3)$$

where $w = [w_{i1}, w_{i2}, \dots, w_{im}]^T$ is the the input nodes. $\beta = [\beta_{i1}, \beta_{i2}, \dots, \beta_{i\tilde{N}}]^T$ represents weighted output that keeps hidden nodes linked to output nodes, and b_i is the bias of the hidden layer. The integral is denoted by the value $w_i \cdot x_q$, and $O = [o_{j1}, o_{j2}, \dots, o_{jN}]^T$ signifies the output vector which is $m \times 1$.

A standard SLFN has \tilde{N} hidden and it can estimate n distinct training examples with zero error, i.e.,

$$\sum_{i=1}^N \|o_q - t_q\| = 0, \quad (4)$$

then there must exist w_i, b_i and β_i which satisfy the function:

$$\sum_{i=1}^{\tilde{N}} \beta_i g_i(x_q) = \sum_{i=1}^{\tilde{N}} \beta_i g_i(w_i x_q + b_i) = t_j, j = 1, \dots, N. \quad (5)$$

The above equation can be written as

$$H\beta = T, \quad (6)$$

where $H(w_1, w_2, \dots, w_i, b_1, b_2, \dots, b_i, x_1, x_2, \dots, x_i,)$

$$= \begin{bmatrix} g(w_1 x_1 + b_1) & \dots & g(w_n x_1 + b_n) \\ \dots & \dots & \dots \\ g(w_1 x_N + b_1) & \dots & g(w_n x_N + b_n) \end{bmatrix}, \quad (7)$$

where H signifies the hidden layer output matrix of network. To reach the optimal minima, the hidden node parameters are adjusted traditionally. In contrast, the output weights of an SLFN can be determined analytically, and we may choose the hidden node parameters with any nonzero activation function. Given this, the estimate procedure for determining the output matrix in terms of the theory of least squares β can be written as

$$\hat{\beta} = H^\dagger T, \quad (8)$$

where H^\dagger is sometimes called the Moore–Penrose generalized inverse, or simply the generalized inverse of H .

The objective of the RELM method is to locate the best answer that will allow the following equation to be satisfied:

$$\|H\tilde{\beta} - T\|_F^2 = \min_{\beta} \|H\beta - T\|_F^2, \quad (9)$$

where $\|\cdot\|_F$ is known as Frobenius norm. Several regularization methods have been published in the literature, including minimax concave, ridge regression, and nonconvex term. These notations have been used to lessen the standard deviation of linear systems. It has been shown that ELM begins to overfit the model when the sum of hidden nodes is more than 5000. Recognizing that the linear system representing the output of these SLFNs can be determined analytically, we employ the Frobenius norm as a means of regularization. The equivalent form of (9) is as follows:

$$\|H\tilde{\beta} - T\|_F^2 = \min_{\beta} (\|H\beta - T\|_F^2 + \lambda \|\beta\|_F^2), \quad (10)$$

$$\tilde{\beta} = (H^T H + \lambda I)^{-1} H^T T. \quad (11)$$

For regularized ELM, $\tilde{\beta}$ is calculated as exposed in Eq. (11) where λ is a factor used for regularization. Equation (11) provides the best answer to Eq. (10) when λ is a positive constant term (i.e., > 0). Adjusting the ratio of empirical risk to structural risk is achieved through manipulating. A generalized model can be developed by making the optimal trade-off between these two risks. This optimization process is carried out by RFO, which is described as follows:

3.4.1 Overview of Red Fox Optimization (RFO)

The red fox is an efficient predator of rodents, birds, and small mammals. Some red foxes stick to well delineated territory, while others are migratory [26]. Under the rule of the alpha couple, each herd is responsible for a certain area. When the young one reaches the adulthood, they may choose to leave the herd and establish their own if they believe they would have a better chance of establishing dominance in a new region. Fox hunting area is typically passed down through generations or remains in the family.

3.4.1.1 Fundamental Principle Apiece population is considered by an n -point $\bar{x} = (x_0, x_1, \dots, x_{n-1})$. In the symbolization (\bar{x}_j^i) is population and j is the organize to recognize each fox x^i in iteration t . Let $f \in R^n$ be the normal charm; $(\bar{x})^{(i)} = [(x_0)^{(i)}, (x_1)^{(i)}, \dots, (x_{n-1})^{(i)}]$ denotes the dimensions. Each space is specified as $(a, b)^n, a, b \in R$. If the value of

function $f(\bar{x}^{(i)})$ is a global value on $(\bar{x}^{(i)})$, this is the perfect solution (a, b) .

3.4.1.2 Global Exploration Phase Each fox in a pack is essential to the group's continued existence. When food is scarce, or just out of curiosity, individual members of the herd may go to more distant places. Therefore, we first sort the population by fitness condition, and then we use Eq. (12) to get the square of the Euclidean distance to each member of the population as $(\bar{x}^{\text{best}})^t$.

$$\text{dis}\left((\bar{x}^{(i)})^t, (\bar{x}^{\text{best}})^t\right) = \sqrt{(\bar{x}^{(i)})^t - (\bar{x}^{\text{best}})^t}. \quad (12)$$

Transferring members of a population in the most beneficial direction is modeled by the Eq. (13).

$$(\bar{x}^{(i)})^t = (\bar{x}^{(i)})^t + \alpha \text{sign}\left((\bar{x}^{\text{best}})^t - (\bar{x}^{(i)})^t\right). \quad (13)$$

For all entities in the population, $\alpha \in \text{dist}\left((\bar{x}^{(i)})^t, (\bar{x}^{\text{best}})^t\right)$ is α indiscriminately chosen, climbing iterative total.

3.4.1.3 Traversing Done the Local Habitat-Local Search Stage The arbitrary value $\mu \in (0, 1)$ was previously calibrated to simulate the fox's probability of being spotted as it approaches its prey, defined by the Eq. (14)

$$\begin{cases} \text{Move closer if } \mu > 0.75 \\ \text{Stay and disguise if } \mu \leq 0.75 \end{cases} \quad (14)$$

This iteration of demonstrates restraint in population transfer while using an enhanced Cochleoid equation to illustrate the separate effort. For this effort, the scaling restriction $a \in (0, 0.2)$ is randomly selected for all memberships of a population to signify different places, from the target to the fox's arrival, and the initial viewing angle $\varphi_0 \in (0, 2\pi)$ is selected for all animals to mimic the fox's perspective. It aids in the calculation of the vision radius of a foraging fox, which is given by the Eq. (15).

$$R = \begin{cases} a \frac{\sin \varphi_0}{\varphi_0} & \text{if } \varphi_0 \neq 0 \\ \theta & \text{if } \varphi_0 = 0 \end{cases}, \quad (15)$$

where θ is a random value among 0 and 1. Model actions of the scheme spatial organizes are stated in Eq. (16):

$$\begin{cases} x_0^{\text{new}} = ar \cdot \cos(\varphi_1) + x_0^{\text{actual}} \\ x_1^{\text{new}} = ar \cdot \sin(\varphi_1) + ar \cdot \cos(\varphi_2) + x_2^{\text{actual}} \\ x_2^{\text{new}} = ar \cdot \sin(\varphi_1) + ar \cdot \sin(\varphi_2) + ar \cdot \cos(\varphi_3) + x_2^{\text{actual}} \\ \dots \dots \\ x_{n-2}^{\text{new}} = ar \cdot \sum_{k=1}^{n-2} \sin(\varphi_k) + ar \cdot \cos(\varphi_{n-1}) + x_{n-2}^{\text{actual}} \\ x_{n-1}^{\text{new}} = ar \cdot \sin(\varphi_1) + ar \cdot \sin(\varphi_2) + \dots + ar \cdot \sin(\varphi_{n-1}) + x_{n-1}^{\text{actual}} \end{cases} \quad (16)$$

Every angular valuation is randomized, discussing to $\varphi_1, \varphi_2, \dots, \varphi_{n-1} \in (0, 2\pi)$.

3.4.1.4 Reproduction Stage Two of the strongest characters must die off every so often to maintain the total number of people at a stable $(\bar{x}^{(1)})^t$ and $(\bar{x}^{(2)})^t$ are chosen, to indicate the alpha link, Eq. (17) computes the habitat centre:

$$\text{habitat}^{(\text{center})^t} = \frac{(\bar{x}^{(1)})^t + (\bar{x}^{(2)})^t}{2}. \quad (17)$$

Distance between the specified parameters, expressed as the square of the distance, is given by Eq. (18):

$$\text{habitat}^{(\text{diameter})^t} = \sqrt{\|(\bar{x}^{(1)})^t - (\bar{x}^{(2)})^t\|}. \quad (18)$$

The aforesaid distances between the stated parameters are illustrated by Eq. (18). The ideal solution in a given number of iterations is a reversal of all point classifications that are constrained by random values. Iteratively, the function is represented in terms of the distance among the alpha functions. Random constraints, expressed as $k \in (0, 1)$ in terms of Eq. (19), are taken into account at each iteration.

$$\begin{cases} \text{New nomadic individual if } k \geq 0.45 \\ \text{Reproduction of the alpha couple if } k < 0.45 \end{cases} \quad (19)$$

If copying an optimal key for two personalities, $(\bar{x}^{(1)})^t$ and $(\bar{x}^{(2)})^t$ join in a new separate $(\bar{x}^{\text{reproduced}})^t$, as shown in Eq. (20):

$$(\bar{x}^{\text{reproduced}})^t = k \frac{(\bar{x}^{(1)})^t - (\bar{x}^{(2)})^t}{2}. \quad (20)$$

In this article, the study examines RFO's temporal complexity, where n is the population size, D is the dimension of the issue, and T is the total number of iterations. All entities are sorted out in each cycle, yielding a function of complexity $O(n \times D)^2$. In particular for high-dimensional applications, RFO's temporal complexity can be reduced by employing a fast-sorting procedure. The computing time is defined by ignoring the low-order components that may be present, as $O(3 \times T \times n^2 \times D^2)$ functions.

Algorithm 1 Red fox optimization procedure

```

1: Start,
2: Define parameters of the algorithm: fitness function  $f(\cdot)$ , size of
  search space solution  $\langle a, b \rangle$ , number of iterations  $T$ , the maximum size of the
  population  $n$ , fox observation angle  $\phi_0$ , weather conditions  $\theta$ ,
3: Generate population consisting of  $n$  foxes at random within search
  space,
4:  $t := 0$ ,
5: while  $t \leq T$  do
6: Define coefficients for iteration: fox approaching change  $a$ ,
  scaling parameter  $\alpha$ ,
7: for each fox in current population do
8: Sort individuals according to the fitness function,
9: Select  $(x_{best})t$ ,
10: Calculate reallocation of individuals according to Eq. (13),
11: if reallocation is better than the previous position then
12: Move the fox,
13: else
14: Return the fox to previous position,
15: end if
16: Choose parameter  $\mu$  value to define noticing the hunting fox,
17: if fox is not noticed then
18: Calculate fox observation radius  $r$  according to Eq. (15),
19: Calculate reallocation according to Eq. (16),
20: else
21: Fox stays at his position to disguise,
22: end if
23: end for
24: Sort population according to the fitness function,
25: Worst foxes leave the herd or get killed by hunters,
26: New foxes are replaced in the population using Eq. (19) as a
  nomadic fox outside the habitat or are reproduced from the alpha
  couple inside the herd Eq. (20),
27:  $t++$ ,
28: end while
29: Return the fittest fox  $(x)_{best}$ ,
30: Stop.

```

4 Results and Discussion

The study employed accuracy, mean ROC curve score, mean precision, mean recall, and mean F1 score to evaluate the performance of the different designs. A multi-label classifier's prediction may only be accurate to some extent. To calculate precise match accuracy, the study considers only accurate predictions that exactly match the label vector across all dimensions, ignoring the partially correct ones. The following Eq. (21) provides the formula for the same:

$$\text{Accuracy} = \frac{\text{exactly matched instances}}{\text{total instances}}. \quad (21)$$

Area under the ROC curve indicates the likelihood that a classifier will assign a higher rank to a random positive occurrence than a random negative instance. All labels assigned by each classifier are averaged and their ROC-AUCs are reported:

$$\text{ROCAUC}_{\text{Mean}} = \frac{1}{L} \sum_{i=1}^L \text{ROCAUC}(l_i). \quad (22)$$

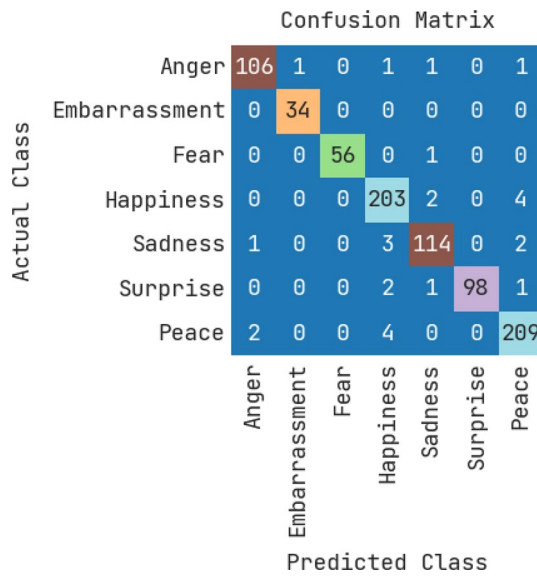


Fig. 6 Confusion matrix for training results

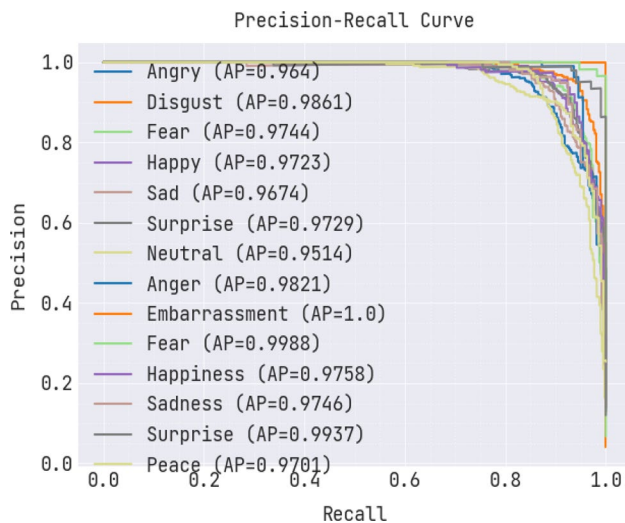


Fig. 7 Precision-recall curve for various emotions

The accuracy of a prediction is measured by how many times it was right out of all the times it was forecasted as right. The study's average accuracy is calculated by averaging the Y-axis label precisions.

$$\text{Precision}_{\text{Mean}} = \frac{1}{L} \sum_{i=1}^L \frac{|Y_i \cap Z_i|}{|Z_i|}. \quad (23)$$

The recall metric measures how many accurate examples were predicted out of all the correct cases. In this analysis, we utilise a measure of memory called “average recall,” which is calculated as the weighted average of recall for each label.

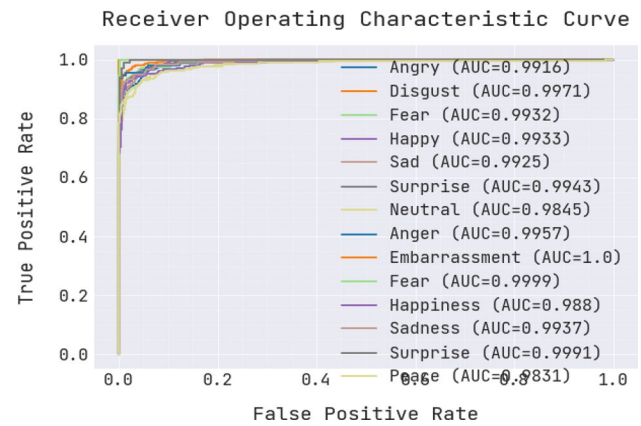


Fig. 8 ROC for training data

$$\text{Recall}_{\text{Mean}} = \frac{1}{L} \sum_{i=1}^L \frac{|Y_i \cap Z_i|}{|Y_i|}. \quad (24)$$

$F1$ score is the harmonic mean of precision and recall and the average $F1$ score is given by:

$$F1_{\text{Mean}} = \frac{1}{L} \sum_{i=1}^L \frac{2|Y_i \cap Z_i|}{|Y_i| + |Z_i|}. \quad (25)$$

4.1 Analysis of Proposed Model in EMOTIC Dataset

An extensive analysis of the results obtained for EMOTIC dataset¹ is provided in this section.

As observed in Fig. 6, the proposed model, for training data accurately identified ‘Anger’ in 106 instances with minor misclassifications into ‘Embarrassment’, ‘Sadness’, and ‘Peace’. ‘Embarrassment’ was perfectly classified 34 times, and ‘Fear’ was accurately predicted in 56 instances with no errors. For ‘Happiness’, the model made 203 correct predictions but misclassified it as ‘Sadness’ and ‘Peace’ a few times. ‘Sadness’ was mostly identified correctly (114 instances) but had minor misclassifications involving ‘Anger’, ‘Happiness’, ‘Surprise’, and ‘Peace’. ‘Surprise’ saw 98 accurate predictions with slight deviations with ‘Happiness’ and ‘Peace’. Lastly, the model performed perfectly for ‘Peace’ with 209 correct identifications, though it deviated from optimal performance with ‘Happiness’, ‘Sadness’, and ‘Surprise’. Overall, the proposed model exhibited satisfactory performance in emotion classification, with certain areas, especially between ‘Happiness’, ‘Sadness’, and ‘Peace’, presenting opportunities for fine tuning.

In Fig. 7, each emotion’s Average Precision (AP) value, a summary metric of the precision-recall curve, signifies the

¹ <https://www.kaggle.com/datasets/magdawjcicka/emotic/data>.

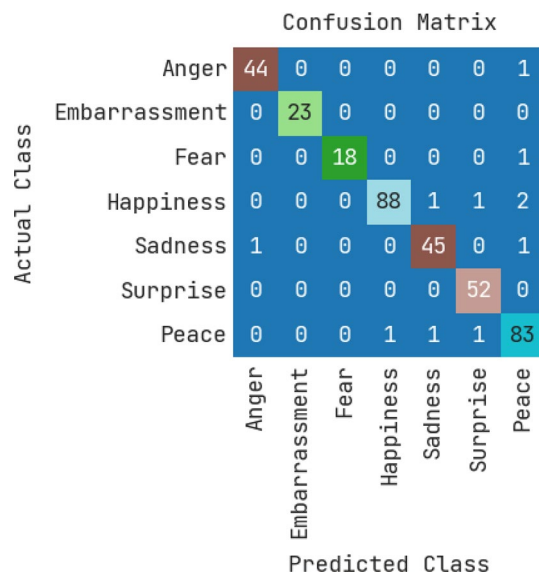


Fig. 9 Confusion matrix for testing data

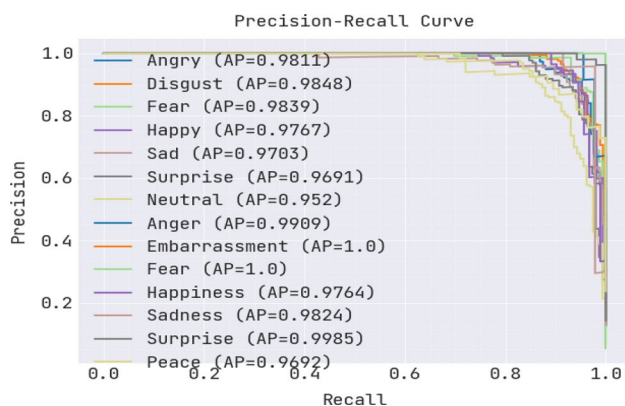


Fig. 10 Precision-recall curve for testing data

model's efficacy. With emotions like "Angry" at AP 0.9641, "Disgust" at 0.9861, "Fear" at 0.9784, "Happy" at 0.9723, "Sad" at 0.9764, "Surprise" at 0.9729, "Neutral" at 0.9514, another instance of "Anger" at 0.9211, "Embarrassment" at a perfect 1.0, and "Peace" at 0.9872, the model consistently exhibits high precision across different recall levels for the training set.

Figure 8 presents the Receiver Operating Characteristic (ROC) curves for diverse emotions, mapping the True Positive Rate against the False Positive Rate. The associated Area Under the Curve (AUC) values signify the model's classification efficacy for each emotion. Notably, "Angry" achieves an AUC of 0.9916, "Disgust" at 0.9971, "Fear" at 0.9990, "Happy" at 0.9953, "Sad" at 0.9925, "Surprise" at 0.9943, "Neutral" at 0.9845, another variant of "Anger" at

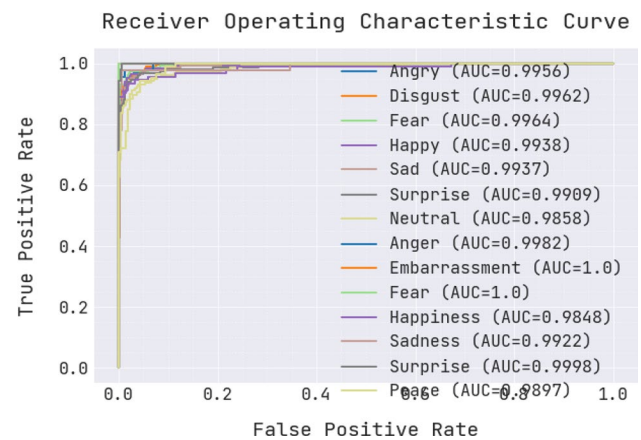


Fig. 11 ROC curve for testing data

Table 2 Analysis comparison of proposed model

Metrics	Testing	Training
Accuracy	0.9698	0.9681
Precision	0.9757	0.974
Recall	0.972	0.9715
F1-score	0.9737	0.9726
ROC AUC score	0.995	0.9942

0.9957, "Embarrassment" with a perfect 1.0, and "Peace" at 0.9831. Overall, the model demonstrates significant emotion classification performance which can be evidenced from the positioning of the curves which is towards the top-left corner, indicating high AUC values.

For training data, EMOTIC dataset is divided into 70% of data and results are taken for precision-recall curve with ROC in terms of different emotions such as angry, fear, happy, sad, disgust, neutral, surprise and embarrassment. The remaining 30% data of EMOTIC is considered as testing data.

As observed in Fig. 9, the proposed model, for test data, exhibited accurate classifications which are evident from the following observations: "Anger" with 44, "Embarrassment" 23, "Fear" 18, "Happiness" 88, "Sadness" 45, "Surprise" 52, and "Peace" 83. However, slight misclassifications are noted, such as "Happiness" being mistaken for "Sadness" and "Surprise" or "Anger" for "Peace." Overall, the model demonstrates commendable accuracy in emotion detection, with minor overlaps primarily between "Happiness," "Sadness," and "Surprise."

In Fig. 10, the precision-recall curves provide notable AP values which include "Angry" with 0.9811, "Disgust" at 0.9848, "Fear" reaching 0.9830, "Happy" achieving 0.9767, and "Sad" at 0.9763. Other emotions such as "Surprise,"

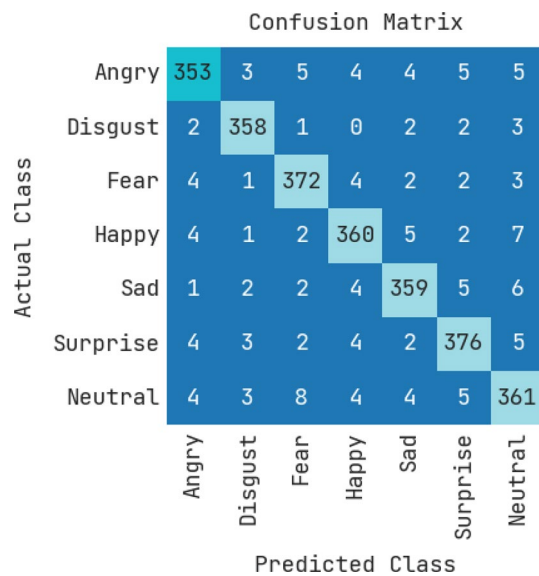


Fig. 12 Confusion matrix for training data

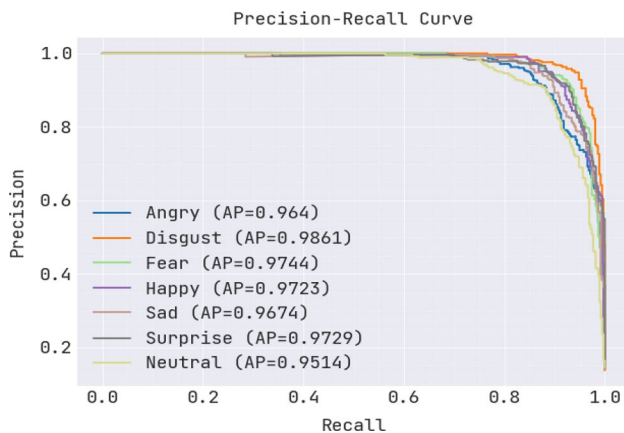


Fig. 13 Precision–recall for different emotions

“Neutral,” and a second representation of “Anger” also display high AP values of 0.9691, 0.9952, and 0.9890 respectively. Remarkably, “Embarrassment” stands out with a perfect AP of 1.0. In contrast, “Peace” has an AP of 0.9822. Collectively, these curves and AP values suggest a robust model performance across the depicted emotions.

As seen in Fig. 11, each emotion’s Area Under the Curve (AUC) score reflects the classifier’s effectiveness. Emotions like “Angry,” “Disgust,” “Fear,” “Happy,” “Sad,” “Surprise,” and “Neutral” display commendable AUC values, ranging from 0.9956 to 0.9988, indicating excellent model performance. Remarkably, “Embarrassment” and “Fear” both achieve a perfect AUC score of 1.0, representing flawless

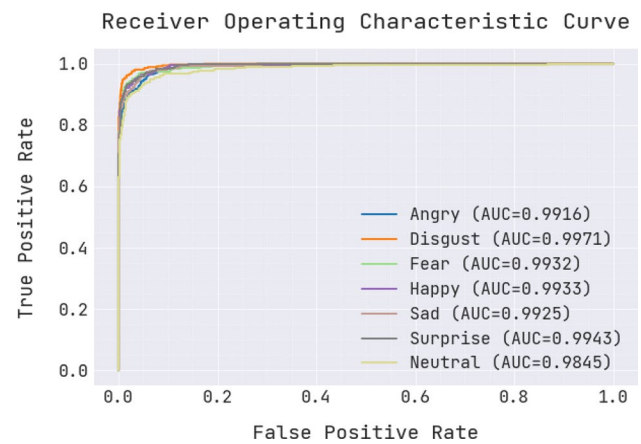


Fig. 14 ROC curve for training data with various emotions

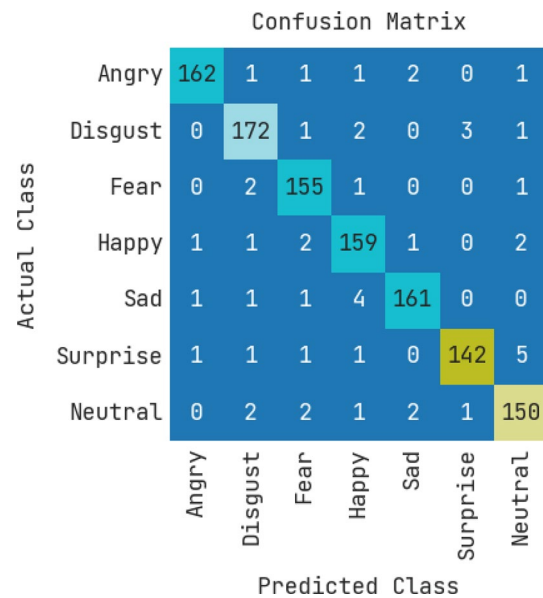


Fig. 15 Confusion matrix for testing data of FER2013

classification. “Peace,” with an AUC of 0.8997, suggests slightly lower performance compared to other emotions. Overall, the depicted ROC curves and corresponding AUC values denote a proficient model in distinguishing between various emotional states.

Table 2 provides the different metrics of proposed model for training and testing data.

In the analysis of accuracy, the proposed model achieved 96% for training and testing data, proposed model has 99% of AUC score for both data. When testing with precision, recall and F1-score, the proposed model achieved 97% for training data and 97% of testing data.

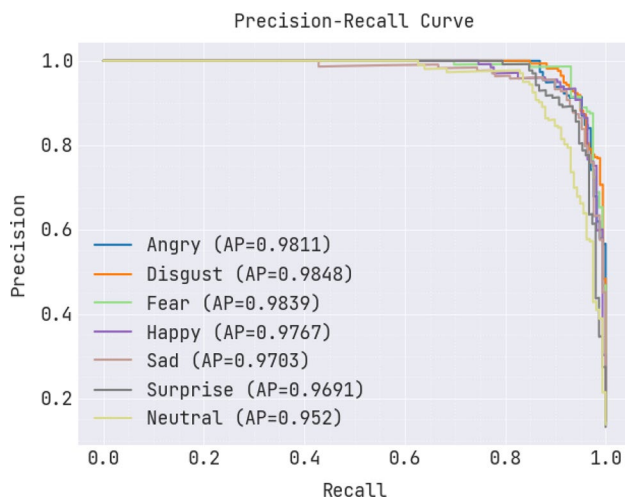


Fig. 16 Precision-recall curve for testing data

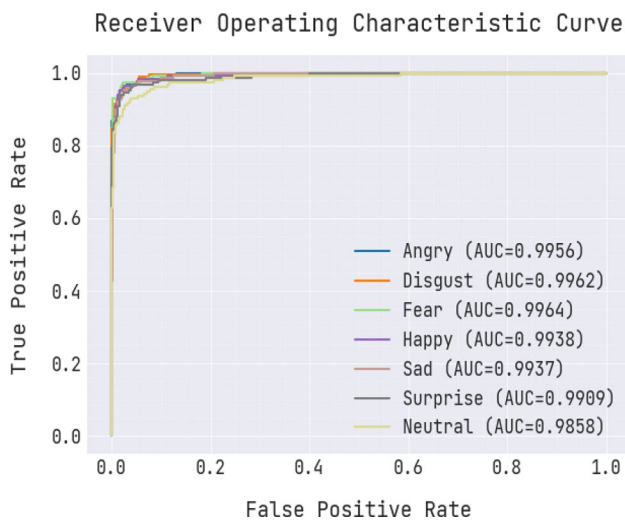


Fig. 17 ROC Curve of testing data with various emotions

4.2 Analysis of Proposed Model in FER2013 Dataset

An extensive analysis of the results obtained for FER2013 dataset² is provided in this section.

Figure 12 presents the confusion matrix for the results obtained for FER2013 dataset. As observed, diagonal values, like 353 for “Angry” and 358 for “Disgust,” represent correct predictions, indicating substantial accuracy for these emotions. However, some off-diagonal values signify misclassifications, for instance, “Happy” being mistaken as “Sad” 5 times or “Surprise” misclassified as “Neutral” 5 times. Among all emotions, “Fear” demonstrates notable accuracy

Table 3 Performance analysis of proposed model on FER2013 dataset

Metrics	Testing	Training
Accuracy	0.9582	0.9474
Precision	0.9584	0.9475
Recall	0.958	0.9475
F1-score	0.9581	0.9475
ROC AUC score	0.9932	0.9924

Table 4 Comparative analysis of proposed model with existing techniques

Author/year	Dataset	Accuracy
Bodavarapu, P.N.R. [14] 2021	FER2013	95
Sivaiah, B., [15] 2022	FER2013	94
Bendjoudi, I., [20] 2021	EMOTIC	95
Athavle, Madhuri et.al [27] 2021	FER2013 dataset	95.00
Dar et.al [28] 2021	FER2013	93.40
Nixon [29] 2022	FER2013	82.0
Agarwal et.al [30] 2022	EMOTIC dataset	96.40
Proposed	EMOTIC dataset	96.98
Proposed	FER2013	95.82

The bold indicates the results of the proposed model against the two datasets

with 372 correct predictions and minimal errors across other emotions.

As observed in Fig. 13, most curves closely approach the top right corner, indicating high precision and recall values. Specifically, emotions like “Angry” and “Disgust” have Average Precision (AP) scores of 0.964 and 0.9861, respectively, showcasing their superior predictive accuracy. While all emotions achieve commendable AP values exceeding 0.95, “Neutral” has the lowest AP at 0.9514. Overall, the model demonstrates robust performance in classifying most emotional states.

As observed in Fig. 14, Receiver Operating Characteristic (ROC) Curve illustrates the classification model’s performance across different thresholds indicates the proximity of most curves to the top left corner thus justifying high true positive rates with minimal false positives. Specifically, emotions like “Angry” and “Disgust” exhibit Area Under the Curve (AUC) values of 0.9916 and 0.9971, respectively, denoting excellent classification. While all emotions present AUC values cross beyond 0.9, “Sad” demonstrates the lowest AUC at 0.9253, indicating potential for enhanced precision in its categorization.

Compared with EMOTIC dataset, the FER2013 dataset has only few emotions such as angry, disgust, fear, happy, sad, surprise and neutral. Here also, the training data is splits into 70% and remaining data of FER2013 is splits into 30%.

² <https://www.kaggle.com/datasets/msmbare/fer2013>.

Table 5 Ablation analysis of proposed model in terms of accuracy

Metric	Proposed FER model (without optimization)	Proposed FER model (with optimization)
Accuracy	0.9451	0.9582

As shown in Fig. 15, for the test case, diagonal entries display accurate predictions, with emotions such as “Angry” and “Disgust” showing high accuracies of 162 and 172, respectively. However, off-diagonal numbers indicate misclassifications. For example, the “Happy” class has been mistakenly identified as “Fear” twice. Similarly, the emotion “Surprise” was misclassified as “Neutral” five times.

As shown in Fig. 16, “Angry” possesses the highest AP at 0.9811, suggesting exceptional model performance for this emotion. However, “Neutral” has the lowest AP at 0.952, indicating potential areas for enhancement. Despite variations, all emotions register high AP values, confirming the model’s general efficacy in predicting emotions accurately.

In the ROC analysis for test data, depicted in Fig. 17, Emotions like “Angry” and “Disgust” have high AUC values of 0.9956 and 0.9962, respectively, indicating superior model performance for these categories. On the other hand, “Neutral” has an AUC of 0.9858, which, although lower than some, still signifies a commendable classification ability. Overall, the curve and its accompanying AUC values reflect the model’s robust performance across different emotions.

Table 3 presents the performance analysis of proposed model for various metrics on FER2013 dataset.

The above table proves that the proposed model achieved 95% of accuracy on testing data and 94% of accuracy on training data. When the proposed model is verified with training data, it achieved 94% of precision, recall and F1-score and the same model achieved 95% of precision, recall and F1-score on testing data. In the analysis of ROC, the proposed model achieved 99% for both training and testing data.

4.3 Comparative Analysis of Proposed Model with Existing Techniques

Table 4 presents the comparative analysis of proposed model with existing techniques for two publicly available datasets in terms of accuracy.

All existing techniques are implemented with both FER2013 and EMOTIC datasets and results are tabulated. From this analysis, it is clearly proved that the proposed model achieved 95.82% of accuracy on FER2013 and 96.98% of accuracy on EMOTIC dataset. The existing techniques mostly achieved 94–95% of FER and 96.40% of

accuracy on EMOTIC dataset. The existing techniques such as DenseNet [14] and CapsNet [15] did not use any optimization models for improving the classification accuracy. In comparison to existing methods for the FER2013 dataset, the proposed model demonstrated approximately a 1% higher accuracy performance than the models presented by Bodavarapu [14] and Athavle [27]. Moreover, when compared to the research models of Sivaiah [15], and Dar et al. [28], the proposed model exhibited a superior accuracy of around 2%. Specifically, the proposed model showcased a 13% improvement in accuracy compared to the Nixon [29] model for the FER2013 dataset. Likewise, for the EMOTIC dataset, the proposed model displayed an approximately 2% higher accuracy than Bendjouidi [20], and a 1% better accuracy than Agarwal et al. [30] research models. The reason for better performance is that the features are extracted by using DL-based feature extraction and classification is carried out by using WRELM, where kernels are optimized by using RFO.

Additionally, the accuracy of the FER model can be investigated in terms of ablation result (i.e., with and without optimization) in order to derive the superiority of the proposed model. It can be seen from Table 5 clearly that the proposed FER model with RFO optimization yielded a higher accuracy of 0.9582 compared to that of the FER model without optimization that yielded an accuracy of 0.9451.

5 Conclusion

In this research work, the face detection is carried out by using MTCNN models and then, normalization with equalization process is carried out for effective enhanced images. From the pre-processed image, features are extracted by using EfficientNetB0 model by removing the unwanted features, filter sizes are changed by using only convoluted features. Therefore, the computation time and power are minimized. Finally, the classification is done by weighted KELM and RFO is used to select the optimal kernel parameters to improve the classification accuracy. Two datasets such as EMOTIC and FER2013 are used to test the effectiveness of proposed model in terms of accuracy, precision, recall and F1-score. The analysis proves that the proposed model achieved 95.82% of accuracy on FER2013 and 96.98% of accuracy on EMOTIC dataset. However, the proposed model focused on binary emotion classification. The EMOTIC dataset has a greater number of emotions, which may be focused on future studies by improving the classification model with feature selection techniques. Advances in neural networks and computational capabilities are expected to significantly improve the accuracy and efficiency of emotion recognition algorithms. One possible avenue is the integration of these

models into various sectors, ranging from healthcare for patient monitoring to customer service for gauging customer satisfaction in real-time. There is also the prospect of blending other types of data with facial expressions, such as vocal intonations or physiological signals, to create more comprehensive emotion recognition systems. Also, enhancement of human–computer interaction, where computers could adapt or respond based on the user’s emotional state.

Author Contributions MA: research proposal, construction of the work flow and model, final drafting, survey of existing works and improvisation of the proposed model; SB: initial drafting of the paper, collection of datasets and choice of their suitability and formulation of pseudocode.

Funding The author did not receive support from any organization for the submitted work.

Availability of Data and Materials The author hereby declare that no specific data sets are utilized in the proposed work. The have also agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Code Availability Since, future works are based on the custom codes developed in this work, the code may not be available from the author.

Declarations

Conflict of Interest The author has no relevant financial or non-financial interests to disclose.

Ethical Approval The paper is an original contribution of research and is not published elsewhere in any form or language.

Consent Statement All authors mentioned have contributed towards the research work, drafting of the paper as well as have given consent for publishing of this article.

Consent to Publication All authors listed above have consented to get their data and image published.

Humans or Animals Participant No Humans or Animals were involved in the experimentation.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Assed, M.M., Khafif, T.C., Belizario, G.O., Fatorelli, R., Rocca, C.C.D.A., de Pádua Serafim, A.: Facial emotion recognition in maltreated children: a systematic review. *J. Child Fam. Stud.* **29**(5), 1493–1509 (2020)
2. Manoharan, J.S.: Design of an intelligent approach on capsule networks to detect forged images. *J. Trend. Comput. Sci. Smart Technol.* **3**(3), 205–221 (2021)
3. Mehendale, N.: Facial emotion recognition using convolutional neural networks (FERC). *SN Appl. Sci.* **2**(3), 1–8 (2020)
4. Alreshidi, A., Ullah, M.: Facial emotion recognition using hybrid features. *Informatics* **7**(1), 1–8 (2020)
5. Samuel Manoharan, J.: A smart image processing algorithm for text recognition, information extraction and vocalization for the visually challenged. *J. Innov. Image Process.* **1**(1), 30–38 (2019)
6. Li, K., Jin, Y., Akram, M.W., Han, R., Chen, J.: Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy. *Vis. Comput.* **36**(2), 391–404 (2020)
7. Akhand, M.A.H., Roy, S., Siddique, N., Kamal, M.A.S., Shimamura, T.: Facial emotion recognition using transfer learning in the deep CNN. *Electronics* **10**(9), 1036–1043 (2021)
8. Liu, X., Cheng, X., Lee, K.: GA-SVM-based facial emotion recognition using facial geometric features. *IEEE Sens. J.* **21**(10), 11532–11542 (2020)
9. Canedo, D., Neves, A.J.: Facial expression recognition using computer vision: a systematic review. *Appl. Sci.* **9**(21), 4678–4685 (2019)
10. Agrawal, A., Mittal, N.: Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. *Vis. Comput.* **36**(2), 405–412 (2020)
11. Verma, G., Verma, H.: Hybrid-deep learning model for emotion recognition using facial expressions. *Rev Socionetwork Strategies* **14**(2), 171–180 (2020)
12. Saxena, A., Khanna, A., Gupta, D.: Emotion recognition and detection methods: a comprehensive survey. *J. Artif. Intell. Syst.* **2**(1), 53–79 (2020)
13. Chowdary, M.K., Nguyen, T.N. and Hemanth, D.J.: Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Comput. Appl.* 1–18 (2021)
14. Bodavarapu, P.N.R., Srinivas, P.S.: An optimized neural network model for facial expression recognition over traditional deep neural networks. *Int. J. Adv. Comput. Sci. Appl.* **12**(7), 5432–5443 (2021)
15. Sivaiah, B., Gopalan, N.P., Mala, C. and Lavanya, S.: FL-CapsNet: facial localization augmented capsule network for human emotion recognition. *Signal Image Video Process.* 1–9 (2022)
16. Liu, H., Cai, H., Lin, Q., Li, X., Xiao, H.: Adaptive multilayer perceptual attention network for facial expression recognition. *IEEE Trans. Circ. Syst. Video Technol.* **32**(9), 6253–6266 (2022)
17. Chaudhari, A., Bhatt, C., Krishna, A., Mazzeo, P.L.: ViTFER: facial emotion recognition with vision transformers. *Appl. Syst. Innov.* **5**(4), 80–91 (2022)
18. Hoang, M.H., Kim, S.H., Yang, H.J., Lee, G.S.: Context-aware emotion recognition based on visual relationship detection. *IEEE Access* **9**, 90465–90474 (2021)
19. Fujisawa, A., Matsumoto, K., Yoshida, M., Kita, K.: Emotion estimation method based on emoticon image features and distributed representations of sentences. *Appl. Sci.* **12**(3), 1256–1264 (2022)
20. Bendjoudi, I., Vanderhaegen, F., Hamad, D., Dornaika, F.: Multi-label, multi-task CNN approach for context-based emotion recognition. *Inf. Fus.* **76**, 422–428 (2021)

21. Yang, J., Lv, Z., Kuang, K., Yang, S., Xiao, L., Tang, Q.: RASN: using attention and sharing affinity features to address sample imbalance in facial expression recognition. *IEEE Access* **10**, 103264–103274 (2022)
22. Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollar, P. and Zitnick, C. L.: Microsoft COCO: common objects in context. *Computer Vision -Lecture Notes in Computer Science*, vol. **8693**, pp. 740–755.
23. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through ade20k dataset. *Int. J. Comput. Vis.* **127**, 302–321 (2019)
24. Soleimanipour, A., Azadbakht, M., Rezaei Asl, A.: Cultivar identification of pistachio nuts in bulk mode through EfficientNet deep learning model. *J. Food Measure. Characteriz.* **16**, 2545–2555 (2022)
25. Li, J., Shi, W., Yang, D.: Fabric wrinkle evaluation model with regularized extreme learning machine based on improved Harris Hawks optimization. *J. Textile Inst.* **113**(2), 199–211 (2022)
26. Połap, D., Woźniak, M.: Red fox optimization algorithm. *Expert Syst. Appl.* **166**, 1–14 (2021)
27. Athavle, M., Mudale, D., Shrivastav, U., Gupta, M.: Music recommendation based on face emotion recognition. *J. Inf. Electric. Electron. Eng. (JIEEE)* **2**(2), 1–11 (2021)
28. Tarim, D., Javed, A., Bourouis, S., Hussein, H.S., Alshazly, H.: Efficient-SwishNet based system for facial emotion recognition. *IEEE Access* **10**, 71311–71328 (2022)
29. Nixon, D., Mallappa, V.V., Petli, V., HosgurMath, S.: A novel AI therapy for depression counseling using face emotion techniques. *Global Trans. Proc.* **3**(1), 190–194 (2022)
30. Agarwal, S., Mukesh Kumar Gupta, D.R.: Emotion detection using context based features using deep learning technique. *J. Theor. Appl. Inf. Technol.* **100**(19), 1–12 (2022)
31. Jin, W., Yu, H. and Luo, X.: CvT-ASSD: convolutional vision-transformer based attentive single shot MultiBox detector. In: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), Washington, DC, USA, pp. 736–744 (2021). <https://doi.org/10.1109/ICTAI52525.2021.00117>
32. Xia, N., Yu, H., Wang, Y., Xuan, J., Luo, X.: DAFS: a domain aware few shot generative model for event detection. *Mach. Learn.* **112**, 1011–1031 (2023)
33. Lin, Y., Xu, C., Yu, H., Tian, P., Luo, X.: Incremental event detection via an improved knowledge distillation-based model. *Neuro-computing* (2023). <https://doi.org/10.1016/j.neucom.2023.126519>
34. <https://www.kaggle.com/datasets/msambare/fer2013>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.