International Journal of Data Informatics and Intelligent Computing (IJDIIC)
Vol. 4, No. 2, June 2025, pp. 1~15
DOI: 10.59461/ijdiic.v4i2.166

Fine-Tuned CNNs with Self-Attention Mechanism for Enhanced Facial Expression Recognition

Rabika Khalid[1], Atta Ur Rahman[1], Sania Ali[2], Bibi Saqia[2]
[1]Riphah Institute of System Engineering (RISE), Riphah International University, Islamabad, 46000, Pakistan
[2]Department of Computer Science, University of Science and Technology, Bannu, 28100, Pakistan

Article Info

Article history:

Received February 05, 2025
Revised March 28, 2025
Accepted April 03, 2025

Keywords:

Facial Emotions Recognition
Emotion recognition
Image analysis
CNN
Facial dynamics

ABSTRACT

The growing need for facial emotion recognition in various domains, particularly in online education, has driven advancements in Artificial Intelligence (AI) and computer vision. Facial expressions are a vital source of nonverbal communication as they convey a wide range of emotions through subtle changes in facial features. Recent developments in Deep Learning (DL) and Convolutional Neural Networks (CNNs) have opened new avenues for analyzing and interpreting human emotions. This study proposes a novel CNN-based real-time facial expression recognition (FER) framework tailored for online education systems. The framework incorporates dynamic region attention and self-attention mechanisms, enabling the model to focus on key facial regions that vary in importance depending on emotional context. The proposed model is fine-tuned to enhance its capability to identify facial expressions in various situations by integrating these methods with transfer learning. Experimental results demonstrate that the model achieves an accuracy of 83% using FER 2013, surpassing traditional static image-based techniques. This study proposes to bridge the gap in facial expression observation in online education, facilitating educators with valuable visions into pupil sentiments to advance learning consequences.

Corresponding Author: Atta Ur Rahman (e-mail: atta.rahman@riphah.edu.pk)

1. INTRODUCTION

The conversion to online education has considerably changed how knowledge is distributed, making it difficult for both students and educators. The main issue and challenges depend on efficiently checking and understanding student commitment in real-time. Communication and emotional intellect are important to generate active, approachable learning environments. However, current online platforms have insufficient tools to measure students' emotional conditions, which are important to perceive their learning development. Responsive signals meaningfully affect assignation, maintenance, and whole academic outcomes, but they are still often overlooked in the simulated learning space. Considering this gap, there is a growing petition for schemes that can perceive and understand emotional indications, mainly facial expressions, to increase the learning experience. The proposed study reports this requirement by suggesting a novel CNN framework for real-time facial FER in online education structures. Facial expressions are an influential and universal means of nonverbal statement, outstanding cultural limitations, and efficient transmission of views, opinions, and emotional states [1]. Various studies recognized seven important facial expressions: anger, disgust, fear, sadness, surprise, pleasure, and neutrality [2]. These expressions form the root of sentiment identification systems, with subsequent exploration by Matsumoto and Heider further authenticating their universality and prominence in conversation [3]. Databases, like the ICML 2013 collection, deliver massive resources for training and testing FER structures, encircling a wide variety of words, including happiness, anger, sadness, fear, and neutral conditions [4].

FER methods have extensive uses in different areas, including user experience, mental health, community production, and safety rules [5]. FER methods can help therapists examine patients' emotional conditions during conduct and assist in detecting circumstances such as anxiety, stress, or depression. For

workplaces, FER could be used to estimate employee confidence and emphasis, leading to improved efficiency and teamwork. In surveillance and security, FER supports detecting doubtful activities or potential dangers in real time, refining community protection in places such as train stations, airports, and public locations. Therefore, while conventional FER models have seen substantial growth, they still face limitations in accurately identifying subtle emotional signals and managing the dynamic real-world environment of emotional words [6]. This issue is particularly evident in applications that require real-time processing, such as online education, where emotional engagement significantly influences the learning process. The existing tools often fail to provide insights necessary for real-time emotional engagement in the context of online education. Instructors frequently lack timely feedback on their students' emotional states, which hinders their ability to adjust teaching strategies for enhanced learning activities. The deficiency of such methods creates hurdles between teachers and students, decreasing the capability for personalized knowledge. Sensory data processing and Real-time imaging, combined with progress in facial detection technology, hold the latent to bridge this gap [7]. The suggested CNN framework is designed to train and organize models competently within online educational surroundings through using approaches such as strategy profile, callbacks, segmentation, and hyperparameter optimization [8].

The purpose of the structure presented in this study is to authorize educators to deliver immediate visions into students' emotional states. This feedback will allow instructors during online classes to adjust their teaching methods, improving student engagement and overall learning outcomes [9]. This technique is particularly relevant given the existing lack of specified tools to observe pupils' facial expressions in daily online learning events, such as exams, video lectures, and communications [10]. The suggested technique shows a robust 83% accuracy rate in real-time FER. The proposed model represents its potential to identify even subtle emotional variations [11][12].

While FER's uses spread beyond teaching into domains such as workplace productivity, security, and healthcare, technology is still evolving. In healthcare applications, FER systems can improve patient observation by perceiving emotional suffering in patients, which is particularly significant for ageing and those getting long-term precautions. Real-time emotion tracking can increase employee well-being and make more productive surroundings in workplaces. FER has important potential in safety applications, as it can assist in detecting distrustful attitudes or threats in public locations. For industries, studying customer sentiments in real time delivers valuable perceptions into consumer manners, eventually refining service delivery and customer fulfilment.

There are no specific techniques presently available for real-time tracking of pupils' facial expressions during online events, such as online exams or video conference lectures, despite the potential of FER. This restraint confuses the effective one-to-one care of engagement and emotional states in computer-generated professional sceneries. Furthermore, current FER models often need important computational resources, making them challenging to deploy in real-time use. The suggested framework lectures these challenges by concentrating on low computational difficulty while preserving maximum accuracy. This assists the model to be used in real-time learning and professional environments without compromising outcomes. The FER plays a significant role in various applications, including human-computer association, psychological analysis, and security observation [13]. DL-based FER models face challenges such as disparities, occlusions, illumination, and subtle appearance variations that remain to barricade accuracy despite notable growth. CNNs have presented astonishing performances in image-based gratitude activities due to their ability to absorb classified features. Therefore, traditional CNNs often struggle with seizing long-range dependencies and contextual relations, which are important for precise FER.

To report these limitations, this work suggests an improved FER framework by assimilating self-attention mechanisms into fine-tuned CNN designs. The self-attention mechanism allows the model to capture total dependencies through facial features, thus refining the recognition of subtle languages. The model assists with pre-trained feature extraction while adjusting to domain-specific FER datasets by leveraging fine-tuned CNNs. This hybrid method increases both feature discrimination and strength against differences in facial expressions. The proposed technique assimilates CNN architectures with dynamic region attention mechanisms, allowing the system to focus on relevant facial regions that change in importance depending on the expressive setting. This technique familiarizes its attention dynamically, leading to further accurate feelings and deciphering contrasting static CNN models that treat all facial counties equally. Particular layers are working to capture both spatial associations and temporal differences in facial expressions, further increasing the model's aptitude to identify emotions precisely. The contributions of this study are as follows:

- The development of a novel Fine-tuned CNNs with a self-attention mechanism for real-time facial expression recognition tailored to online education environments.
- The integration of dynamic region attention mechanisms improves the model's adaptability and accuracy in detecting emotional states.

- The incorporation of advanced techniques like geometric feature extraction and facial landmark detection to enhance the recognition of emotional expressions.
- The creation of a novel system capable of providing real-time feedback to educators, enabling them to adjust teaching strategies based on students' emotional states.

The rest of the paper is structured as follows: Section 2 defines the related work of the related literature, with limitations of previous work and the novelty of the proposed study. Section 3 defines the proposed methodology. Section 4 states the experimental setup of the proposed work. Section 5 states the result and discussion. Section 6 shows the conclusion and recommendations for future works.

2. LITERATURE REVIEW

It is observed that various obstacles exist in the current education system, such as the lack of communication between teachers and students, which limits the capacity to customize learning opportunities to meet the requirements of every student. Adaptive learning systems (ALS) have appeared as a viable method for customizing education by combining AI with data analytics. In [11], several classification algorithms, including logistic regression, Linear discriminant analysis (LDA), k-nearest neighbours algorithm (K-NN), regression trees, Naïve Bayes (NB), and support vector machine (SVM), were used for facial expression recognition. Emotion recognition was performed using deeply learned multi-channel textual and EEG features [14]. The individual's facial expressions disclose numerous details about their mental processes and thoughts. The primary objective of real-time emotion detection is to give the computer a human-like capacity to identify and interpret human emotions. This work developed a model to categorize a face picture into one of the seven emotions under consideration in their study by developing a multi-class classifier [15]. Novel ideas and methodology have been introduced by researchers, leading to substantial progress in the field of facial emotion identification with CNNs. For FER, several approaches have been studied over the past few years [16]. The conventional methods characterize emotions by using values extracted only from facial image attributes. Contemporary DL methodologies incorporate various stages and extract features from the different hierarchical processes [17]. Studies have investigated and compared modern FER methods, namely deep learning-based methods. In [18], the authors presented an impressively accurate human group face sentiment recognition system using CNNs and Haar filters.

The authors of [19] achieved the highest single-network classification accuracy by overcoming FER barriers with VGGNet. Meanwhile, [20] significantly improved over baseline results by using CNN to classify emotions using facial images. The speaker emotion identification was performed via processing approaches such as speech signal, from which features are extracted for final categorization. The speech processing techniques, which are mostly based on periodic and spectral characteristics, provide useful information for emotion classification. In certain cases, voice recognition systems help with categorization by utilizing linguistic information [21]. Various studies have substantially contributed to increasing the models' accuracy and efficiency on the active side of FER. Another face expression recognition study used a densely connected convolutional network. In other studies, remarkable FER test accuracy was achieved by using deep CNNs for image-based facial expression identification. Their results were optimistic because of their capacity to convey the complexity of facial emotions.

The study conducted in [22] focused on model training details to achieve high accuracy, like using CNNs to identify emotions on the face. Additionally, they used pre-trained models to improve the accuracy. The study conducted in [23] developed a new system corresponding to five basic emotional expressions. Their method mimics human facial emotions based on facial skin colour and texture. The Inflated 3D Convolutional Networks for video-based facial emotion recognition were developed, which represented a significant development for facial expression recognition. This method highlights the importance of temporal information for understanding facial emotions, integrates two and three-dimensional convolutional filters, and performs well across various data types [24]. Furthermore, a noticeable trend is observed in using CNN architecture for the prediction of emotions and facial actions in multi-task learning [25]. In another study, researchers have come up with an interesting solution to the problem of face expression mapping because of the challenges posed by the picture's two dimensions and the subsequent digital image analysis using the region of interest. They used lip characteristics to analyze the lips according to different emotions. Their method implied that focusing on several activities at once improves performance as a whole and leads to a deeper comprehension of facial expressions [23]. Finally, the study examining CNN models' performance in recognizing emotions from facial expressions has shown significant progress, with implications for human-computer interaction and affective computing. In [19], authors examined the possibility of using edges to help convolutional neural networks recognize emotions. The research investigated how CNN's architecture can be used to analyze facial emotions, which improves the edge information. The previous studies consistently explore novel designs, datasets, and methods for better performance of their systems. These schemes jointly represent the growth of CNN-based facial expression

recognition. Though a lot of work is done in this field, more work and enhancement are still required to improve facial emotion recognition [26].

2.1. Previous Work Limitation

It is observed that in various studies, the CNN model was used as "black boxes," making it challenging to comprehend how they make predictions. Similarly, various studies ignored certain individual and cultural differences in expression in favour of assuming universal face emotion representations. They also ignored temporal dynamics in emotion categorization [27]. Moreover, the dependence on facial images avoids the sequential nature of feelings and sentiments [28]. The major challenges created by the nonuniformity of the human expression, face, and some extra limitations linked with shadows, facial position, location, and lighting regarding numerous situations [29]. After thoroughly studying the literature, it's mandatory to correctly assume human feelings and sentiments. The conventional approaches mainly lack context-specific features of face reading.

2.2. Proposed work novelty

- We developed a novel approach for FER by using tine-tuned CNNs with the self-attention mechanism. The proposed model focused on discriminative features and expression-relevant facial areas, e.g., eyes and mouth, while suppressing irrelevant features and solving various challenges observed in facial images like occlusions and lighting variations.
- We applied an iterative optimization approach specifically tailored for expression-related characteristics, which improves the model's sensitivity to nuanced facial features. Furthermore, self-attention improves the model adjustment for noisy and obscured pictures, enhancing the model performance for facial expression detection.
- The proposed model incorporates both raw pixel data and geometric facial landmarks (e.g., eyes, eyebrows, mouth contours) as prior knowledge to enhance feature learning. The model prioritized anatomically relevant regions for facial expression analysis.

3. METHODOLOGY

This section explains the proposed methodology of this work, as described in Figure 1. The goal of the proposed study is to develop a self-attentive FER technique leveraging fine-tuned CNNs and a self-attention mechanism. This method attempts to achieve higher identification accuracy by merging the feature extraction abilities of CNNs with the contextual understanding provided via self-attention. CNN model is employed to mine features from it and accomplish the classification once the real-time pictures are fed into the scheme.

3.1. Real-Time Images of Facial Expressions

The facial recognition method works by associating and comparing a picture or video frame of a human face with a record of stored facial images. This method has become progressively predominant with the advent of ID authentication schemes that can perceive and asses distinctive facial features to validate user characteristics. These organizations employ progressive systems to map facial landmarks and extract main features like the nose, position of the eyes, jawline, and mouth. These methods can consistently detect persons across different situations, generating a facial signature. The research emphasizes seven important sentiments in the context of facial expression identification: surprise, sad, happiness, fear, anger, disgust, and neutral. These feelings signify an extensive range of human responses, which can be efficiently perceived using machine learning models trained on different datasets. The aptitude to identify these words precisely in real-time applications holds the possibility for improving collaborating schemes in numerous fields.

3.2. Proposed Application

The suggested method influences CNNs to grow an accurate and proficient FER method. This scheme is precisely considered to handle a wide variety of real-world situations, such as online job interviews, learning sceneries, video conferencing, and virtual business sessions. The classifier's ability to comprehend and accomplish facial expressions robustly permits a united combination of these usages, growing user associations and allowing better communication. For example, in educational situations, this system can permit schools to display pupils' daily conduct carefully, detect signs of disconnection, misperception, or suffering, and permit educators to take timely actions to help pupils. The application also supports customer conduct analysis in online businesses, permitting businesses to tailor their facilities and products to better suit customer sentiments. The system reaches this by studying facial expressions in real-time, certifying it adjusts to different user demographics and ecological circumstances. A real-life instance of

the anticipated solution for emotion recognition shows its flexibility and significance in increasing both individual and professional activities.

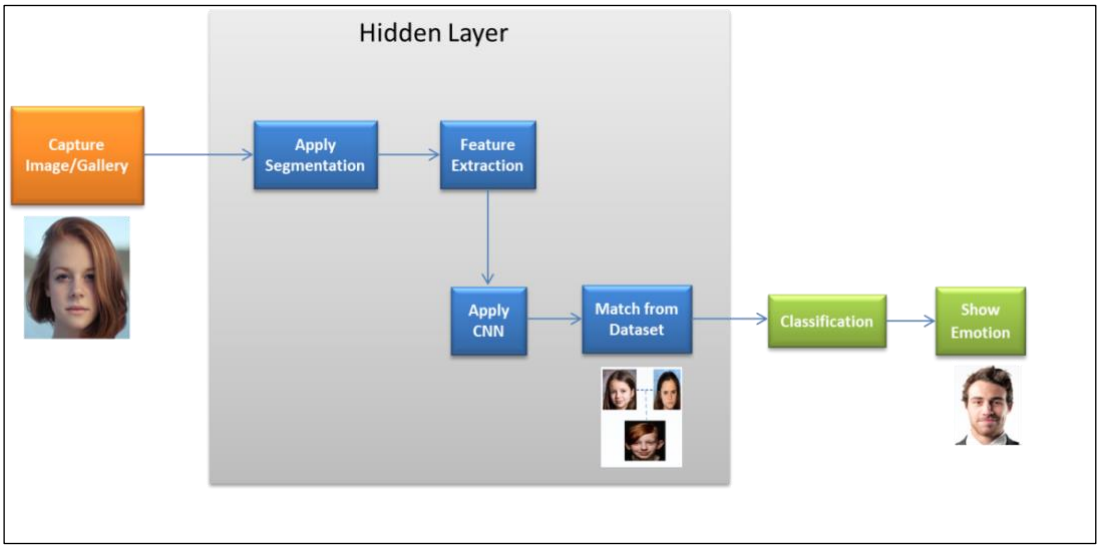

Figure 1. Proposed Methodology

3.3. Image Detection using Landmarks

Facial landmarks are important in increasing the accuracy of FER methods through structural and geometric information regarding the facial areas. In this study, we use facial landmark identification to identify important human face traits. This makes it easier to integrate augmented reality and allows us to keep an eye on head posture, which guarantees driver concentration. These landmarks are usually noticed when pre-trained models recognize important facial themes such as the eyebrows, nose, eyes, jawline, and mouth. The recognized themes are vital for standardizing facial records by justifying the properties of differences in posture, scale, and illumination circumstances. Landmark recognition helps in image preprocessing by assigning and cropping facial areas to a reliable alignment. This step certifies that the input imageries for CNNs emphasize areas critical for expression examination.

The proposed model uses facial landmark identification to clearly address variations in facial muscle movements during expression recognition. By combining the landmarks with a self-attention mechanism within the CNN framework, the proposed model dynamically directs its attention toward regions with high expressive significance, such as mouth, eye, etc. Unlike standard self-attention, which understands primary features purely from raw pixels, our approach utilized landmarks as spatial anchors to bias the attention toward the structurally relevant facial regions. This incorporation of CNN feature extraction and landmark-based position ensures enhanced facial appearance identification. We implement our study using prebuilt libraries such as Media Pipe, OpenCV, and dlib to demonstrate real-world use. The key points used in the study are to identify the location and rotation of a human head posture. Facial recognition involves primary landmarks that serve as key reference points that describe the central structural features, such as the nose tip and the corners of the eyebrows, mouth, and eyes. These landmarks convey a basis for understanding the geometry of the facial expression. Meanwhile, secondary landmarks' performance as secondary points improves and refines the accuracy of the primary landmarks. These processes assist in capturing advanced characteristics, such as contours, relative positioning, and curves. The total landmarks identify the complete set of primary and secondary points that work together to generate a thorough facial structure. Utilizing the landmark features, the proposed model accomplishes greater precision in emotion detection and facial recognition. The details of the landmarks are given in Table 1.

3.3.1. Dataset Preprocessing

As mentioned in the previous section, we used the FER-2013 dataset [30] for the proposed study. The dataset used for training and evaluation was first preprocessed. The proposed algorithm is trained using the FER-2013 dataset, which has thousands of images for every emotion. The classifier used was CNN, consisting of several layers, each carrying out a distinct change in the architecture and learning process. Figure 2 shows the detailed workflow of the proposed work. Table 2 shows the statistics of the FER-2013 dataset. Table 3 contains the statistics of the images we used to train the proposed model against each emotion.

Table 1. Primary and secondary landmarks

Primary landmarks		Secondary landmarks	
No.	**Exactness**	**Total**	**Description**
1	Left eyebrow outer-restrict	63,64	Eye centers
2	Left eyebrow inner-restrict	1	Left sanctuary
3	Right eyebrow inner-restrict	9	Chin overturn
4	Right eyebrow outer- restrict	2-7,11-16	Cheek delineations
5	Left eye outer –restrict	17	Right sanctuary
6	Left eye inner –restrict	19-21	Left eyebrow delineations
7	Right eye inner-restrict	24-26	Right eyebrow curves
8	Right eye outer-restrict	30,31	Nose saddles
9	Nose tip	32,36	Nose peaks (nostrils)
10	Left mouth corner	28-31	Nose contours
11	Right mouth corner	50-54, 56-60	Mouth contours

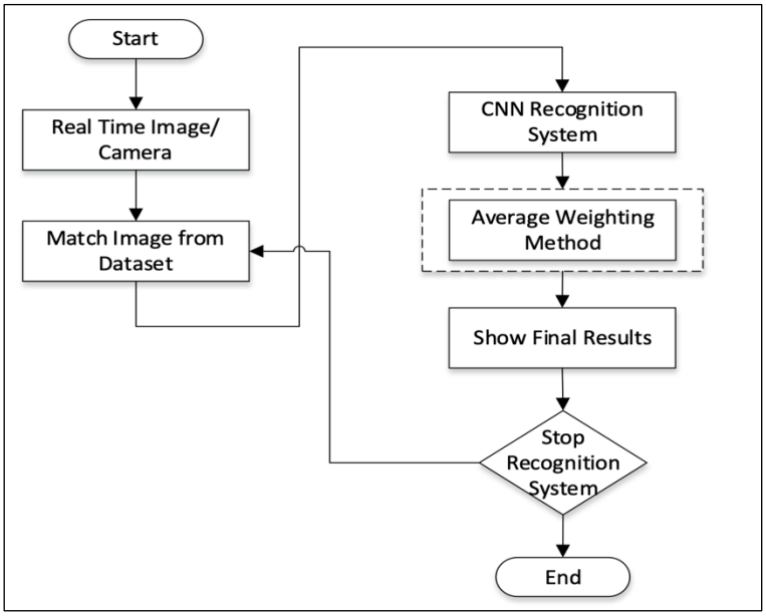

Figure 2. Workflow of the proposed study

3.3.2. Emotion Recognition using CNN

The convolution layer uses various filters to extract spatially localized features such as edges, textures, and expression-specific patterns like frown lines or smile curves from the input images. A pooling layer is used to decrease the computational cost and mitigate overfitting while preserving only discriminative regions. After that, the flattening layer is used to convert the 3D feature maps into a 1D vector for classification. The fully connected (FC) layer is then used to map the obtained features into classes such as happy, angry, etc. The two main parts of the proposed model architecture are self-attentive mechanisms and CNNs. The self-attentive method dynamically concentrates on informative regions within facial expressions, while CNNs constitute the backbone of feature extraction. This section outlines CNN's architecture; an input 2D convolutional layer (with 32 filters) and a 2D Max Pooling layer are used to couple three pairs of 2D convolutional layers, each with 64, 128, and 256 filters, with a 2D MaxPooling layer.

The architecture is made up of various layers that carry out distinct functions. The input layer manages the complexity of the data, while convolutions use least squares to determine the local characteristics of the images. Different filters make functions like edge recognition and sharpening possible. In order to enable non-linear feature learning and allow the model to capture complex patterns in facial expressions, the Rectified Linear Unit (ReLU) activation function is used. As the number of features increases in convolution layers, the pooling layers are used to decrease the number of features and maintain only discriminative features. Finally, the vectorized matrix output (1D feature vector) is processed by a fully

connected layer for classification. The seven emotions which are studied in this work are reflected in Figure 3.

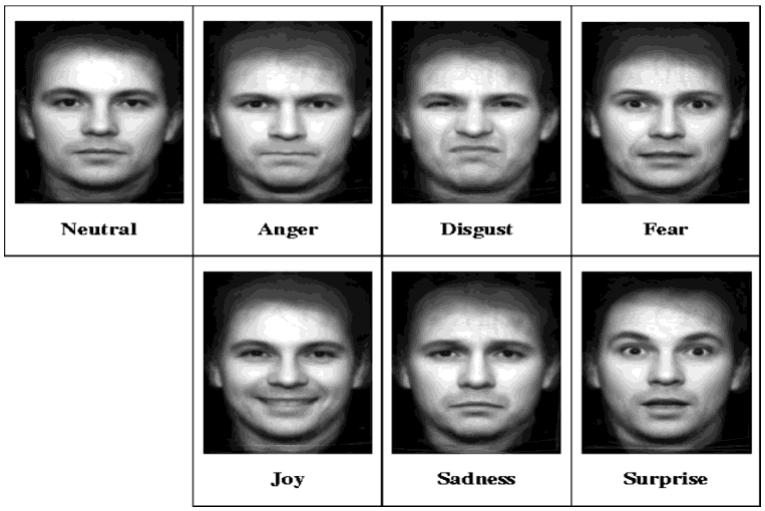

Figure 3. Images showing the seven emotions

Table 2. FER-2013 dataset statistics

Attribute	Description
Image size	48x48 pixels, 35887 images
Task	Facial expression classification
Color format	Grayscale
Face alignment	Centered and similarly scaled
Number of classes	7 (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral)
Class labels	0 = Angry, 1 = Disgust, 2 = Fear, 3 = Happy, 4 = Sad, 5 = Surprise, 6 = Neutral
Training set size	28,709 images
Testing set size	3,589 images
Validation set size	3,589 images

Table 3. Emotion-base dataset statistics used in this work

Sr. No	Name of Emotion	Number of Images
1	Happiness	8989
2	Fear	5121
3	Surprise	4002
4	Sadness	6077
5	Disgust	547
6	Anger	4953
7	Natural	6198

Convolutional layer: A collection of filters, or kernels, are used across the input picture in a convolutional layer to identify the key features. Here, the input image is represented as I and the filter is represented as K. The convolution operation at position (i,j) is described in Eq.1.

$$C(i,j) = \sum m \sum n I(m,n) \cdot K(i-m, j-n) \quad (1)$$

Whereas the above equation computes the value of $C(i,j)$ at a particular location (i,j) in the given image $I(m,n)$ · with kernel $K(i-m, j-n)$. The filters are convolved with the image by calculating the dot product between the kernel weights and each local region of the input image, resulting in a feature map in which each pixel represents the sum of these element-wise products.

Activation function: To add non-linearity and allow the model to learn more complex features, the ReLU activation function is used, which is calculated as in Eq. 2.

$$R(i,j) = \max\big(0, C(i,j)\big) \tag{2}$$

The output of ReLU activation is represented by $R(i,j)$, where the value $\max\big(0, C(i,j)\big)$. If $C(i,j)$ is positive, then the values are retained, removing negative values.

Pooling layer: In order to reduce the spatial dimensions (width & height) of feature maps while retaining the important features, max pooling is used. It decreases the spatial dimensions while maintaining the key features. The pooling operation is described in Eq. 3.

$$P(x,y) = max\big(R(2x,2y), R(2x+1,2y), \quad R(2x,2y+1), R(2x+1,2y+1)\big) \tag{3}$$

Where the maximum value from a 2×2 area of the activation map $R(i,j)$ is retained, this helps to decrease the feature size while maintaining the most significant facial information.

Fully connected layer: The operation in the fully connected layer is given through indicated in Eq. 4.

$$O_k = \text{ReLU}(\sum jWkj \cdot Hj + bk) \tag{4}$$

Here, O_k is the output for the k-th neuron, W_{kj} represents the weights connecting the jth neuron in the previous layer to the k-th neuron in the current layer, and b_k represents the bias term for the k-th neuron.

Output layer: The output layer is used to make predictions based on the processed inputs from previous layers. For multi-class classification, we use softmax activation as calculated in Eq. 5.

$$Softmax(z)_i = \frac{e^{z_i}}{\sum_j e^{z_j}} \tag{5}$$

Where z_i is the input to the softmax layer for class i, and the output is a probability distribution for each class.

Loss function: Due to multi-class classification tasks, the categorical cross-entropy loss function is used, which is calculated as shown in Eq.6.

$$\mathcal{L} = \sum_{i=1}^{k} y_i \log(p_i) \tag{6}$$

Here, y_i represents the true label (one-hot encoded), and p_i shows the predicted probability for class i.

3.3.3. Self-Attentive Mechanism

The self-attention mechanism is used to compute the attention scores across both (channel and spatial dimensions), enhancing the overall feature extraction. For the identification of face expressions, we used a self-attention method to determine the most discriminative regions in the face structures. Using dynamic weighting approaches, various areas of the input image are considered and weighted. This procedure allows the model to concentrate on semantically significant information for the categorization of emotions from facial images. This method was implemented using learnable attention weights and layered attention layers. Equation 7 is used to compute the attention weights.

$$A = \sigma\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{7}$$

Where A represents the attention matrix to find the rank of different spatial areas in the feature map. Q is quary matrix of transform version of input attributes employed to compute the emphasized features.

Meanwhile, K is the change in the input features that helps calculate similarity with enquiries. The purpose of the key matrix is represented by K^T employed to identify relevancy between different locations. The dimensionality of key vectors is denoted by d_k, and the softmax function is represented by σ.

4. EXPERIMENTS

This section discusses the experimental setup, training process, and assessment of the proposed work in detail.

4.1. Training process

The dataset used in this study is based on the FER-2013 facial expression recognition dataset, which contains 35,887 grayscale images of size 48×48 pixels. The dataset is divided into 80% (28709 samples) for training and 20% (3589 samples) for testing the proposed model. The training procedure for the proposed fine-tuned CNNs with a self-attention mechanism begins with preprocessing using the FER-2013 dataset. During this phase, all images are processed on the Roboflow platform using logical management techniques. Enhancements such as rotation, blurring, saturation adjustment, and flipping are applied to improve the quality of the images. Once preprocessing is complete, the training phase begins, with the model achieving a maximum accuracy of 80% after 100 epochs. In the experiment, various iterations are performed. When the model does not reach at least 80% accuracy on the validation set after full training, which is 100 epochs in this work, the results are considered inadequate. It was considered possible problems with the architectural design, hyperparameters, or quality of the dataset. The design and hyperparameters were tuned to obtain optimal results. Throughout training, the deep CNN is optimized to learn from the training data, minimize loss functions, and enhance overall performance. This process relies on several factors, including hyperparameters, callbacks, training duration, and model complexity. Two key callbacks are employed to ensure efficient training. The first is the use of early stopping, which monitors validation accuracy and halts training if the improvement is not greater than 0.00005 within 11 consecutive epochs, restoring the model to the best weights to prevent overfitting. The second callback, ReduceLROnPlateau, reduces the learning rate by a factor of 0.5 if validation accuracy does not improve within 7 epochs. The model's performance is evaluated using accuracy and loss as key metrics.

4.2. Experimental Setup

This section outlines the tools and techniques employed in the experimental procedure. Pre-trained CNN architectures, ResNet-50, are fine-tuned on the preprocessed datasets to leverage their pre-trained features while adapting them to facial expression recognition tasks.

Table 4. Model parameters for fine-tuned CNNs with self-attention framework

Parameter	Description
Dataset	FER-2013
Input Image Sizes	48 × 48 pixels
Pre-trained CNN Model	ResNet-50, InceptionV3
Attention Mechanism	Self-attention layers applied to high-expression regions (e.g., eyes, mouth)
Learning Rate	0.0001 (adaptive scheduling with a reduction on plateau)
Batch Size	32
Number of Epochs	50 (with early stopping based on validation accuracy)
Optimization Algorithm	Adam optimizer
Loss Function	Categorical Cross-Entropy
Regularization	Dropout (rate: 0.5) and L2 regularization
Data Augmentation	Random rotation, horizontal flipping, and brightness
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score, and Confusion Matrix
Facial Landmark Detection	Dlib library for aligning faces and identifying key regions
Hardware Configuration	GPU: NVIDIA GeForce RTX 3090, CPU: Intel Core i9, RAM: 32 GB
Programming Frameworks	TensorFlow and Keras, with NumPy, Pandas, and OpenCV
Model Input	Aligned and preprocessed facial images (RGB)
Model Output	Probabilistic scores for facial expressions

A self-attention mechanism is integrated into the architecture to improve the model's focus on expression-critical regions, such as the mouth, eyes, and eyebrows. To minimize validation loss, the Adam optimizer is utilized with an initial learning rate of 0.0001, combined with an adaptive learning rate scheduler. A batch size of 32 is used to balance memory efficiency and training stability, and the model is trained for 100 epochs with early stopping criteria based on validation accuracy. Dropout regularization and L2 regularization were both applied at a rate of 0.5 to mitigate overfitting. The optimization process ensures that the model effectively learns to distinguish between various facial expression categories. An exponential linear unit (ELU) activation function is used to address the vanishing gradient problem. The training setup includes a system with 32 GB RAM, a Core i9 processor, GPU support, and 256 GB storage. During the training process, associated facial images and detected landmarks guide the self-attention mechanism to prioritize expression-relevant features. The model's performance is evaluated using robust metrics to ensure reliable and consistent recognition of facial expressions. Table 4 presents the model parameters used in the proposed facial expression recognition framework.

4.3. Evaluation Criteria

This section demonstrates the evaluation criteria adopted during the validation of the proposed model to verify the results. Precision, recall, F1-score, and accuracy are used to evaluate the suggested model's efficiency. Precision assesses the model's ability to recognize facial expressions; however, improper positives are reduced. Eq.8 is used to compute it.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

The recall (true positive amount) of the model activates its competence to classify all real facial expression actions while reducing false negatives. The formula in Eq.9 is used to accomplish the computation.

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

TP identifies true positives, FP shows false positives, and FN indicates false negatives. The F1 score computes the harmonic mean of precision and recall and provides a balanced statistic for assessing the model performance, especially in class-imbalanced situations. The Eq. 10 is used to calculate the F1-score.

$$F1 - Sore = \frac{2.Precision.Recall}{Precision + Recall} \tag{10}$$

Accuracy is considered as the ratio of properly predictable cases to total cases in the dataset. Eq.11 is used to perform the calculation.

$$Accuracy = \frac{No\ of\ Correct\ Predictions}{Total\ Nof\ Predictions} \tag{11}$$

5. RESULTS AND DISCUSSION

This section provides a detailed discussion of the results of the proposed method, along with different related studies using advanced DL techniques. The study published in [3] employed CNNs and used the FER-2013 dataset; they obtained a test accuracy of 75.2% without utilizing supplementary training data or face recognition. The work accomplished by [31] developed a method for facial expression recognition. Their proposed approach is attractive, especially considering the success of CNNs in addressing facial recognition problems. Their focus was on achieving high accuracy using a limited training dataset. The CNN utilized a dense Scale-Invariant Feature Transform (SIFT) with the FER-2013 dataset, outperforming traditional CNN methods. They observed that when their models were combined, accuracy significantly improved, achieving better results with an accuracy of 73.4%. The research published by [11] examined the advancement of FER using various DL models, including AlexNet, GoogleNet, and ResNet. Additionally, they identified the key contributions of CNNs to facial expression analysis using the FER-2013 dataset. They explored differences between various DL approaches on the FER-2013 dataset, achieving a maximum accuracy of 64.24% and highlighting the strengths and weaknesses of their work.

The study conducted in [32] employed RCL-Net, a method for detecting wild facial expressions that local binary pattern (LBP) feature extraction and influence attention procedures. The structure contains two main subdivisions: the local binary pattern (LBP) extraction branch and the ResNet-CBAM residual attention branch. First, they link a hybrid attention mechanism with an advanced system, where the residual attention network enhances the local face features more competently. They perceived that it made a robust residual

attention framework through mining significant visual features from both channel dimensions and the spatial. Besides, after facial expression feature extraction, an improved residual attention network was advanced by assimilating the LBP attributes. This integration enhanced the feature representation and increased the identification accuracy by capturing fine-grained texture features in photos of facial expressions. The FER-2013 dataset was used in their studies, and they obtained a higher accuracy of 74.23%. The research published in [18] applied DL techniques to recognize feelings and emotions in crowds of people. They initially use a Haar filter to perceive expressions and extract features. After that, a CNN was employed to recognize facial expressions and categorize them into five sentiments: angry, surprised, happy, sad, and neutral. Finally, the identified sentiments were adapted into audio production using a synthesizer. Their proposed model achieved 65% accuracy on the FER-2013 dataset. The study conducted by [19] attained higher accuracy for a particular system using the FER-2013 dataset. They employed the VGGNet model, using different optimization approaches. Their proposed method achieved an accuracy of 73.28% without using any further training data on FER-2013. The research published in [33] investigated whether a CNN-based model performs better when trained solely on raw pixel data from images.

Table 5. Evaluation of the Proposed Model

Activity	Dividing pictures (%)	Precision (%)	Loss (%)	Error Rate	Error Rate (%)
Train	90	94	53		
Test	5	80	94	9/46	19.56
Validation	5	84	66		
Train	80	94	54		
Test	10	78	79	19/91	20.87
Validation	10	89	66		
Train	70	93	53		
Test	15	78	79	28/100	28
Validation	15	79	79		

Table 6. Accuracies Comparison of Different Techniques on the FER-2013 Dataset

Reference	Framework	Accuracy (%)
[3]	Ensemble of modern deep CNNs.	75.2
[31]	Convolutional Neural Networks	73.4
[11]	CNN, VGGNet and AlexNet.	64.24
[32]	Attention Mechanism and LBP	74.23
[18]	Convolutional Neural Networks	65
[19]	CNN, VGGNet architecture	73.28
[33]	Convolutional Neural Networks	75.1
[26]	Deep Convolutional Neural Network	70
Proposed Study	Fine-Tuned CNNs with Self-Attention Mechanism	83

To enhance the performance, the authors supplemented the raw pixel input using additional features, including Histogram of Oriented Gradients (HOG) and facial landmarks. First, the faces were detected using an LBP classifier, followed by facial landmark extraction using dlib. The HOG features were then computed and integrated into their model. The CNN was trained using both raw pixel data and these supplementary features obtained through LBP and HOG. Their experimental results on the FER-2013 dataset demonstrated an improved accuracy of 75.1%. The study conducted in [26] concentrated on generating a scheme that can identify not just the seven mutual facial terms but also four supplementary ones: "Tired/Exhausted," "Pain," "Showing Interest", and "Lack of Interest" during real-time video calls. The scheme could classify 11 facial terms. They built two CNN models to accomplish this classification of facial expression. Their first model, which they called CNNM9, could categorize faces into nine emotions: "Sad," "Fear," "Happy," "Anger,"

"Disgust," "Pain," "Surprise," "Tired/Exhausted," and "Neutral." Their proposed scheme achieved an accuracy of nearly 78%. Table 5 represents an evaluation of the proposed model. Table 6 shows the comparison of different techniques on the FER-2013 Dataset. Our proposed model achieved an accuracy of 83 % on the FER-2013 dataset using the fine-tuned CNN, along with landmark detection and self-attention mechanism. The proposed work outperformed the state-of-the-art techniques for the task of facial expression recognition. Figure 4 represents the loss of the proposed model.

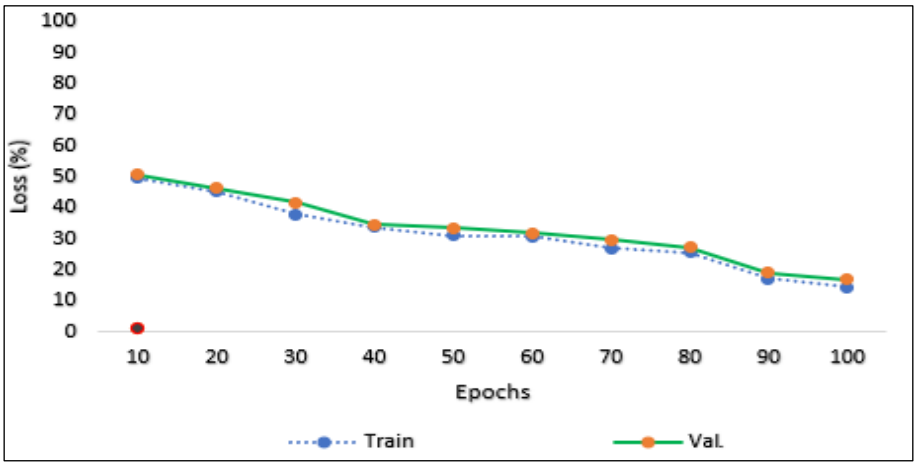

Figure 4. Loss of propose

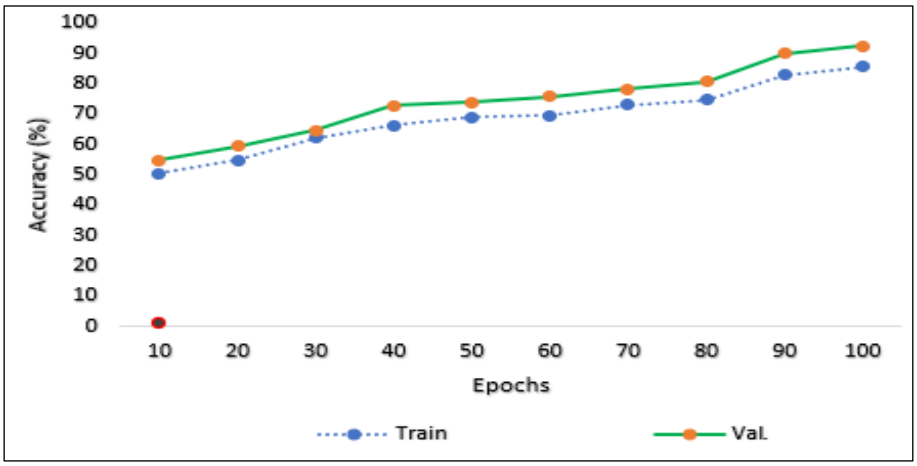

Figure 5. Accuracy of the proposed model

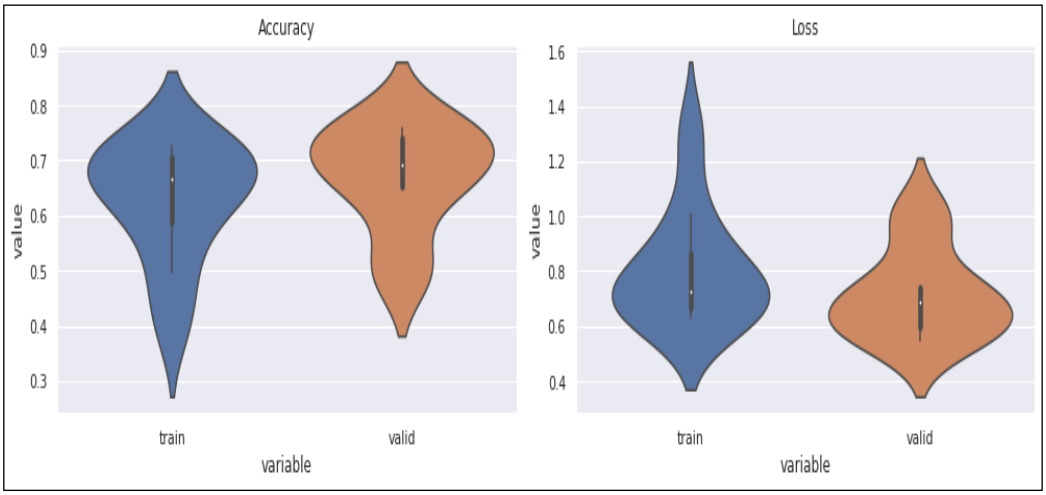

Figure 6. Performance distribution: accuracy (Left) and loss (Right

Figure 5 shows the accuracy of the proposed model. The performance distribution graph is shown in Figure 6, where accuracy is on the left and loss is displayed on the right. For this purpose, the CNN technology, along with landmark detection and self-attention approach, has been implemented using the FER-2013 dataset. The proposed model performed the best results on the validation dataset by effectively training the model and fine-tuning its hyperparameters. It shows proficiency in categorizing neutral emotions. Comparative research shows that the model performs slightly better and provides insights into online learning. The proposed approach improves facial appearance identification by dynamically assigning higher attention scores to key areas of the facemask. This process enables the model to emphasize features such as eye openness, eyebrow movements, and mouth shape variations, which are vital for distinguishing words, expressions, and appearances. Self-attention enhances feature extraction by utilizing global contextual data, ensuring that subtle yet significant patterns are effectively captured and thereby boosting classification results. Figure 7 illustrates the outcomes of the proposed model compared to baseline techniques, indicating that the suggested model surpasses state-of-the-art methods. The subsequent graph demonstrates that the proposed model achieved superior results compared to baseline studies in the same field.

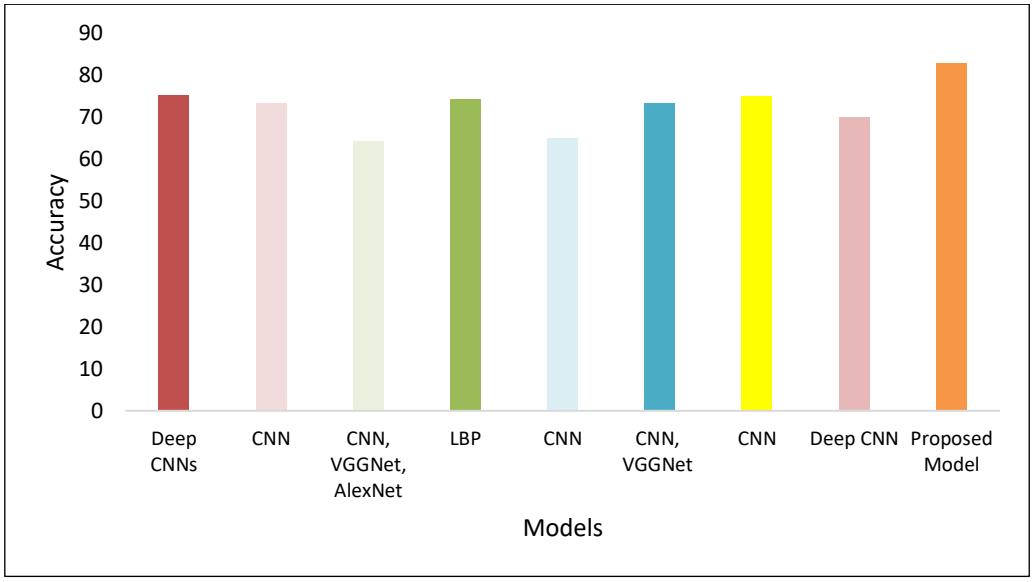

Figure 7. Proposed model comparison with baseline approaches

6. CONCLUSION

The study concludes by introducing a self-attentive technique for improved facial expression identification employing convolutional neural networks integrated with dynamic area attention mechanisms. After a careful review of the literature and extensive testing, we have demonstrated the superiority and efficacy of the proposed methodology in facial expression identification. It was observed that the proposed model achieved better results than the conventional static CNN architectures. This research employed CNN architecture for FER to tackle the issue of real-time student attention detection in online learning. The proposed model demonstrates potential in mental health care, education, and user experience improvement, as it achieved an accuracy of 83% on the validation dataset for facial expression identification. The ability of the proposed model can be utilized to precisely interpret human emotions for general well-being, personalize services, and optimize user experiences. In order to increase model efficiency and interpretability in facial emotion recognition, future research may focus on improving self-attention processes.

Future research will explore the integration of CNN-based emotion recognition systems into online education platforms for real-time student engagement tracking. This implementation will enhance the learning outcomes through an adaptive feedback approach while demonstrating the model's scalability for diverse educational applications.

DATA AVAILABILITY STATEMENT

The data presented in this study are available on request from the corresponding author.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest in this work.

REFERENCES

[1] E. Pranav, S. Kamal, C. Satheesh Chandran, and M. H. Supriya, "Facial Emotion Recognition Using Deep Convolutional Neural Network," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, Mar. 2020, pp. 317–320. doi: 10.1109/ICACCS48705.2020.9074302.

[2] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: review and insights," *Procedia Comput. Sci.*, vol. 175, pp. 689–694, 2020, doi: 10.1016/j.procs.2020.07.101.

[3] C. Pramerdorfer and M. Kampel, "Facial Expression Recognition using Convolutional Neural Networks: State of the Art," Dec. 2016. http://arxiv.org/abs/1612.02903

[4] K. Sarvakar, R. Senkamalavalli, S. Raghavendra, J. Santosh Kumar, R. Manjunath, and S. Jaiswal, "Facial emotion recognition using convolutional neural networks," *Mater. Today Proc.*, vol. 80, pp. 3560–3564, 2023, doi: 10.1016/j.matpr.2021.07.297.

[5] M. M. Taghi Zadeh, M. Imani, and B. Majidi, "Fast Facial emotion recognition Using Convolutional Neural Networks and Gabor Filters," in *2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)*, IEEE, Feb. 2019, pp. 577–581. doi: 10.1109/KBEI.2019.8734943.

[6] W. Gu, C. Xiang, Y. V. Venkatesh, D. Huang, and H. Lin, "Facial expression recognition using radial encoding of local Gabor features and classifier synthesis," *Pattern Recognit.*, vol. 45, no. 1, pp. 80–91, Jan. 2012, doi: 10.1016/j.patcog.2011.05.006.

[7] A. J. O'Toole and C. D. Castillo, "Face Recognition by Humans and Machines: Three Fundamental Advances from Deep Learning," *Annu. Rev. Vis. Sci.*, vol. 7, no. 1, pp. 543–570, Sep. 2021, doi: 10.1146/annurev-vision-093019-111701.

[8] H. Y. Patil, A. G. Kothari, and K. M. Bhurchandi, "Expression invariant face recognition using local binary patterns and contourlet transform," *Optik (Stuttg).*, vol. 127, no. 5, pp. 2670–2678, Mar. 2016, doi: 10.1016/j.ijleo.2015.11.187.

[9] George, A., H. Wimme, and C.M. Rebman, "Artifical Intelligence Facial Expression Recognition for Emotion Detection: Performance and Acceptance," *Issues Inf. Syst.*, 2020, doi: 10.48009/4_iis_2020_81-91.

[10] M. Liu, S. Li, S. Shan, and X. Chen, "AU-inspired Deep Networks for Facial Expression Feature Learning," *Neurocomputing*, vol. 159, pp. 126–136, Jul. 2015, doi: 10.1016/j.neucom.2015.02.011.

[11] Y. Gan, "Facial Expression Recognition Using Convolutional Neural Network," in *Proceedings of the 2nd International Conference on Vision, Image and Signal Processing*, New York, NY, USA: ACM, Aug. 2018, pp. 1–5. doi: 10.1145/3271553.3271584.

[12] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017, doi: 10.1109/JSTSP.2017.2764438.

[13] M. Karnati, A. Seal, D. Bhattacharjee, A. Yazidi, and O. Krejcar, "Understanding Deep Learning Techniques for Recognition of Human Emotions Using Facial Expressions: A Comprehensive Survey," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–31, 2023, doi: 10.1109/TIM.2023.3243661.

[14] Y. Liu and G. Fu, "Emotion recognition by deeply learned multi-channel textual and EEG features," *Futur. Gener. Comput. Syst.*, vol. 119, pp. 1–6, Jun. 2021, doi: 10.1016/j.future.2021.01.010.

[15] R. Pathar, A. Adivarekar, A. Mishra, and A. Deshmukh, "Human Emotion Recognition using Convolutional Neural Network in Real Time," in *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, IEEE, Apr. 2019, pp. 1–7. doi: 10.1109/ICIICT1.2019.8741491.

[16] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Emotion recognition using facial expressions," *Procedia Comput. Sci.*, vol. 108, pp. 1175–1184, 2017, doi: 10.1016/j.procs.2017.05.025.

[17] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network," *Sensors*, vol. 21, no. 9, p. 3046, Apr. 2021, doi: 10.3390/s21093046.

[18] K. P and A. T, "Group Facial Emotion Analysis System Using Convolutional Neural Network," in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, IEEE, Jun. 2020, pp. 643–647. doi: 10.1109/ICOEI48184.2020.9143037.

[19] Y. Khaireddin and Z. Chen, "Facial Emotion Recognition: State of the Art Performance on FER2013," May 2022. http://arxiv.org/abs/2105.03588

[20] A. Bhandari and N. R. Pal, "Can edges help convolution neural networks in emotion recognition?," *Neurocomputing*, vol. 433, pp. 162–168, Apr. 2021, doi: 10.1016/j.neucom.2020.12.092.

[21] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artif. Intell. Rev.*, vol. 43, no. 2, pp. 155–177, Feb. 2015, doi: 10.1007/s10462-012-9368-5.

[22] J. Cai, O. Chang, X.-L. Tang, C. Xue, and C. Wei, "Facial Expression Recognition Method Based on Sparse Batch Normalization CNN," in *2018 37th Chinese Control Conference (CCC)*, IEEE, Jul. 2018, pp. 9608–9613. doi: 10.23919/ChiCC.2018.8483567.

[23] M. H. Alkawaz, D. Mohamad, A. H. Basori, and T. Saba, "Blend Shape Interpolation and FACS for Realistic Avatar," *3D Res.*, vol. 6, no. 1, p. 6, Mar. 2015, doi: 10.1007/s13319-015-0038-7.

[24] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1113–1133, Jun. 2015, doi: 10.1109/TPAMI.2014.2366127.

[25] M. Mohammadpour, H. Khaliliardali, S. M. R. Hashemi, and M. M. AlyanNezhadi, "Facial emotion recognition using deep convolutional networks," in *2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, IEEE, Dec. 2017, pp. 0017–0021. doi: 10.1109/KBEI.2017.8324974.

[26] U. F. Ikwuanusi, "Real Time Classification of Facial Expressions for Effective and Intelligent Video Communication," *Electron. Theses Diss.*, p. 1076, 2023.
[27] A. A. Pise *et al.*, "Methods for Facial Expression Recognition with Applications in Challenging Situations," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–17, May 2022, doi: 10.1155/2022/9261438.
[28] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, Jul. 2022, doi: 10.1109/TAFFC.2020.2981446.
[29] M. Long and Y. Zeng, "Detecting Iris Liveness with Batch Normalized Convolutional Neural Network," *Comput. Mater. Contin.*, vol. 58, no. 2, pp. 493–504, 2019, doi: 10.32604/cmc.2019.04378.
[30] I. J. Goodfellow *et al.*, "Challenges in Representation Learning: A Report on Three Machine Learning Contests," 2013, pp. 117–124. doi: 10.1007/978-3-642-42051-1_16.
[31] T. Connie, M. Al-Shabi, W. P. Cheah, and M. Goh, "Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator," 2017, pp. 139–149. doi: 10.1007/978-3-319-69456-6_12.
[32] J. Liao, Y. Lin, T. Ma, S. He, X. Liu, and G. He, "Facial Expression Recognition Methods in the Wild Based on Fusion Feature of Attention Mechanism and LBP," *Sensors*, vol. 23, no. 9, p. 4204, Apr. 2023, doi: 10.3390/s23094204.
[33] V. S. Amal, S. Suresh, and G. Deepa, "Real-Time Emotion Recognition from Facial Expressions Using Convolutional Neural Network with Fer2013 Dataset," 2022, pp. 541–551. doi: 10.1007/978-981-16-3675-2_41.

BIOGRAPHIES OF AUTHORS

Rabika Khalid is currently pursuing a Master of Science degree in Data Science at Riphah International University, Islamabad. With coursework completed, she delves into her thesis under the supervision of Dr. Atta ur Rehman, exploring the realms of machine/deep learning, image processing, and human-computer interaction, driven by a passion for innovative research at the intersection of data science and technology. She can be contacted at email: 44835@students.riphah.edu.pk

Atta Ur Rahman received the BS (Hons) degree in Telecommunication from USTB in 2014 and the MS degree in Computer Science from the same university. He obtained his Ph.D. degree in Computer Science in 2022, from Ghulam Ishaq Khan (GIK) Institute of Engineering Sciences and Technology, Pakistan. Rahman's research interests include data mining, computer vision, human-computer interaction, NLP, and computational intelligence. He has various publications in reputed journals, including IEEE Transactions. He worked as an Assistant Professor at Riphah Institute of System Engineering (RISE), Riphah International University Islamabad, Pakistan. Currently, his postdoc is in progress at the Interdisciplinary Research Centers for Finance and Digital Economy, King Fahd University of Petroleum & Minerals (KFUPM), Dhahran, Saudi Arabia. He has more than 20 publications in various reputed journals and conferences, including IEEE Transactions. His research interest includes Human-computer Interaction, Artificial Intelligence in healthcare, and Federated learning for privacy preservation. He can be contacted at email: atta.rahman@riphah.edu.pk

Sania Ali received her BS degree in Computer Science and MS degree in Computer Science from the University of Science and Technology Bannu in 2014 and 2019, respectively. Her research interests include but are not limited to data mining and the semantic web. Recommendation system, AI in health care and Natural Language Processing (NLP). She can be contacted at email: sania.ali@ustb.edu.pk

Bibi Saqia received the (MSCS) degree in Computer Science from the University of Science and Technology Bannu in 2018. Her PhD is in progress at the University of Science and Technology Bannu. She has more than 12 publications in various reputed journals and conferences, including IEEE Transactions. Her research interests include data mining, machine learning, and artificial intelligence. She can be contacted at email: saqiaktk@ustb.edu.pk