IUST
Iran University of Science and Technology

Iranian Journal of Electrical and Electronic Engineering

Journal Homepage: ijeee.iust.ac.ir

Performance Comparison of Facial Emotion Recognition: Introducing a Model within the Driver Assistance Framework based on Deep Learning with LBP Feature Extraction for In-Vehicle Applications

Ehsan Ghasemi*, Seyyed Mohammad Razavi*(C.A.) and Sajad Mohamadzadeh*

Abstract: This study proposes a descriptor-based approach combined with deep learning, which recognizes facial emotions for safe driving. Paying attention to the driver's facial expressions is crucial to address the increasing road accidents. This project aims to develop a Facial Emotion Recognition (FER) system that monitors the driver's facial expressions to identify emotions and provide instant assistance for safety control. In the initial stage, Viola-Jones face detection was employed to detect the facial region, followed by Butterworth high-pass filtering to enhance the identified region for locating the eye, nose, and mouth regions, using Viola-Jones face detection. Secondly, the Local Binary Patterns (LBP) feature descriptor is utilized to extract features from the identified eye, nose, and mouth regions. Using 3 RGB channels, the extracted features from these three regions are fed into RessNet-50 and EfficientNet deep networks. The outputs of the two deep learning models' classifiers are combined and integrated using two ensemble methods: ensemble maximum voting and ensemble mean. Based on these combining classifier rules, the performance was evaluated on the JAFFE and KMU-FED databases. The experimental results demonstrate that the proposed method can effectively and with higher accuracy than other competitors recognize emotions in the JAFFE and KMU-FED datasets. The novelty and originality of this paper lie in its significant application in the automotive industry. Implementing our proposed method in a system capable of high accuracy and precision can help mitigate numerous driving hazards. Our approach has achieved 99% and 98% accuracy on the JAFFE and KMU-FED databases, respectively. This high level of accuracy, coupled with its practical relevance, underscores the innovative nature of our work.

Keywords: Ensemble deep learning, combination of classifiers, Driver assistant, Face emotion recognition, Local binary pattern.

1 Introduction

PSYCHOLOGIST Watzlawick emphasizes that every state (words, silence, activity) in human interaction has meaning and that communication is essential to building civil society [1]. Communication consists of 7% verbal communication, 38% para-verbal communication (such as tone analysis), and 55% nonverbal communication (such as facial expressions, gestures, and eye contact). Nonverbal communication is crucial to many aspects of our daily lives and the effectiveness of human interaction. There is a greater probability that positive emotions will develop than negative emotions, such as fear, mistrust, etc. Consequently, facial expressions are the primary means by which we communicate our emotions to the external world and interact with others. We assess how others engage with

Iranian Journal of Electrical & Electronic Engineering, 2024.
Paper first received 07 October 2024 and accepted 28 December 2024.
* The authors are with the Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.
E-mail: smrazavi@birjand.ac.ir.
Corresponding Author: Seyyed Mohammad Razavi.

us through them. A computer can automatically recognize facial expressions and emotions. Human-computer interactions must be rich and robust to be effective, so automatic emotion recognition systems, like those that recognize mood, are crucial. In recent years, emotion recognition has gained increasing popularity in human-computer interfaces. It has also found applications in animation, medicine, and security [2].

Due to the rapid growth of interest in autonomous vehicles, advanced driver assistance systems have been developed to ensure the safety of drivers and society. Monitoring driving behavior to detect drivers' health status is crucial, as it prevents driver distraction. There is a rise in road traffic accidents, and the primary cause of most road accidents is driver error, such as distracted driving, aggressive driving, or impaired driving caused by alcohol or substance consumption. It is possible to identify aberrant driving behaviors using eye tracking, blink analysis, head pose estimation, and facial expression recognition [3].

Khalil et al. [4] provide a concise overview of the classification of various road accidents. Advances in sensors, computational technology, communication technology, road safety, and vehicle safety features have been taken to a new level, commonly known as Intelligent Transportation Systems (ITS). As part of this research, a robust facial expression recognition technique has been developed to deal with drivers' changing emotions. Facial expressions include joy, surprise, anger, Sadness, fear, and disgust. As these fundamental emotions indicate a communication signal that does not occur in everyday verbal interactions, capturing these moments is difficult. As a result, when this happens, a powerful message is sent to our surroundings, urging us to take health and safety precautions immediately [5].

Machine learning is a traditional FER approach that uses facial features that can benefit the implementation of real-time embedded systems. Both fast speed and reliable accuracy can be achieved through machine learning techniques. However, optimal performance cannot be guaranteed [6]. The state-of-the-art (SOTA) studies on FER methods extensively use deep neural network (DNN) techniques. DNN models eliminate the feature extraction, enabling high performance [7].

Computer vision has been challenged by the task of facial emotion recognition (FER). In general, FER consists of the following four main modules: the image enhancement module, the face detection module, the feature extraction module, and the classification module. Image processing techniques such as filtering, wavelet transformation, and noise removal enhance images. Face detection is accomplished through pattern-based matching or statistically based models [8].

In the feature extraction process, local and global features of facial regions are captured, including appearance, shape, texture, motion, landmarks, geometry, and so on. Ultimately, feature detection can be conducted through supervised or unsupervised methods. It is possible to classify or cluster features according to a variety of classifiers and clustering approaches [9].

In facial expression recognition, feature extraction is an essential component. Feature extraction methods can be divided into geometry-based feature extraction and appearance-based feature extraction. In the geometry-based approach, the distances between landmark points, the area, and the angles of the constructed triangles in the facial region are used as features. As part of the appearance-based approach, pixel intensity values and their relationships with adjacent pixels are considered features [10].

In this work, feature extraction performs spatial and textural information extraction using the LBP feature descriptor. These features are interconnected and are utilized to overcome limitations in facial expression recognition through transfer learning. Ultimately, human facial emotions are classified by amalgamating into a group classifier [11].

Two standard datasets were used to test the proposed system: the KMU-FED [12] dataset and the Japanese Female Facial Expression (JAFFE) [13] dataset.

Accordingly, the remainder of the paper is organized as follows: Section 2 discusses the status of artistic works in FER, Section 3 presents the proposed system, and Section 4 discusses the empirical analysis of the proposed system. The paper concludes with Section 5, which provides recommendations for future improvement.

Due to the significant importance of facial expression recognition while driving, we are pursuing a method that achieves high accuracy in this domain. Consequently, we utilize appropriate image preprocessing techniques and implement ensemble voting and ensemble mean approaches within the deep neural networks ResNet-50 and EfficientNet. Our experiments on the JAFFE database yielded 97% and 99% accuracy, while the KMU-FED database achieved 95% and 98% accuracy. These results demonstrate the efficacy of our proposed method.

2 Related work

Six basic emotions have been identified: pleasure, fear, disgust, Sadness, anger, and surprise (except neutral). Using this concept, Ekman developed the Facial Action Coding System (FACS), the gold standard of emotion detection research. As a result, neutrality has now been included in most data sets used for emotion detection.

Primary emotions include happy Face, angry Face, disgusted Face, fearful Face, sad Face, surprised Face, and contemptuous Face [14].

A two-stage machine learning approach was used in early studies on emotion detection. First, image features were extracted, and then, in the second stage, classifiers were used to identify emotions [15]. Handcrafted features such as Gabor wavelets, Haar-like features, linear binary pattern (LBP) features, and edge histogram descriptors are often employed to detect facial expressions. The classifier selects an image with the most suitable emotion [16]. The techniques are more effective on specialized datasets but have significant limitations when applied to challenging datasets with greater intra-class diversity [17].

In recent years, multiple companies have made remarkable advances in neural networks, deep learning, image classification, and visual challenges. Khoury [18] demonstrated that CNNs can identify emotions more accurately. Moreover, zero-shot learning was used to model human facial expressions using the Toronto Face Dataset (TFD) and the Cohn-Kanade dataset (CK+). Authors in [19] trained a neural network using deep learning and translated human images into animated faces to construct a model for Facial Expression (FE) from stylized animation characters. Malahosseini proposed a neural network incorporating top pooling layers, convolutional layers, and four initial layers for facial expression recognition [7]. Using the BDBN network, the authors integrate feature selection and classification into a recursive network and illustrate the importance of input from both components in achieving higher accuracy on the CK+ and JAFFE databases.

The authors of [20] proposed a deep CNN algorithm for annotating noise in valid images using crowdsourcing. Their deep convolutional neural network (DCNN) was enhanced by deploying ten annotators to annotate each image for the required accuracy, with ten labels in the dataset and multiple cost functions. Authors in [21] utilized an Incremental Boosting Convolutional Neural Network (IB-CNN) to enhance self-emotion recognition of faces by using more distinct neurons that performed better. The authors developed an Identity-Aware CNN (AI-CNN) using identity-sensitive contrast reduction during learning. In the same vein, the End-to-End Network Architecture was developed as a canonical model of an end-to-end network architecture [22]. The authors of [1] proposed a fast and effective self-correcting mechanism (SCN) as a solution to minimize uncertainty and prevent ambiguous facial representations (resulting from noisy annotations) from overfitting deep networks. As a result of SCN, uncertainty can be mitigated in two different dimensions: (1) by utilizing a self-attention mechanism to weight training instances within small batches with rank adjustments, and (2) by fine-tuning training instances within ensembles with lower rankings [23]. An attention network referred to as a Region Attention Network (RAN) was developed to sufficiently emphasize the positional variable of Facial Expression Recognition and obstruction of the face regions. A recent review on emotion recognition through facial appearance, multi-attention networks for facial state detection, and attention-based networks for facial emotion recognition are examples of relevant studies in this area. Compared to previous works, all those mentioned above have significantly improved emotion recognition. However, none of these works contain a simple method for identifying the essential regions of the Face for emotion recognition [24]. It suggests focusing more on critical facial areas in a new framework, which utilizes neural attention-evolution networks [25].

A texture-based appearance and texture feature descriptor is presented in this study using the Local Binary Pattern (LBP) feature descriptor. This model utilizes deep learning to overcome existing limitations and ensemble classifiers to increase accuracy.

3 Methodology

According to Figure 1, the overall architecture of the proposed FER system consists of the following components: data preprocessing, face detection, feature extraction, fine-tuning in deep learning models, and the ensemble of classifiers.

3.1 Data preprocessing

There is a variety of image sizes in the database. The image sizes are initially modified to maintain the required input size for a pre-trained CNN model and align with the deep learning process. Data augmentation techniques such as rotation and flipping at different angles, inversion, and vertical and horizontal shifting are employed on the images to prevent the networks from being overfitted. The images are also fed into the grayscale network to reduce network processing time. These methods help artificially increase the diversity of the training dataset, thereby allowing the model to generalize unseen data better. By normalizing images, the pixel values of an image are adjusted to fall within a specific scale or distribution. This process can enhance the quality and consistency of images, making them more suitable for subsequent analyses, such as feature extraction or model training. Normalization involves scaling the pixel values to a predetermined range, in this case restricted to [-1, 1]. This adjustment helps make the data more uniform, improving machine learning algorithms' performance. A normalization process is applied to the model dataset to adjust the range of pixel intensity values to a specified extent.

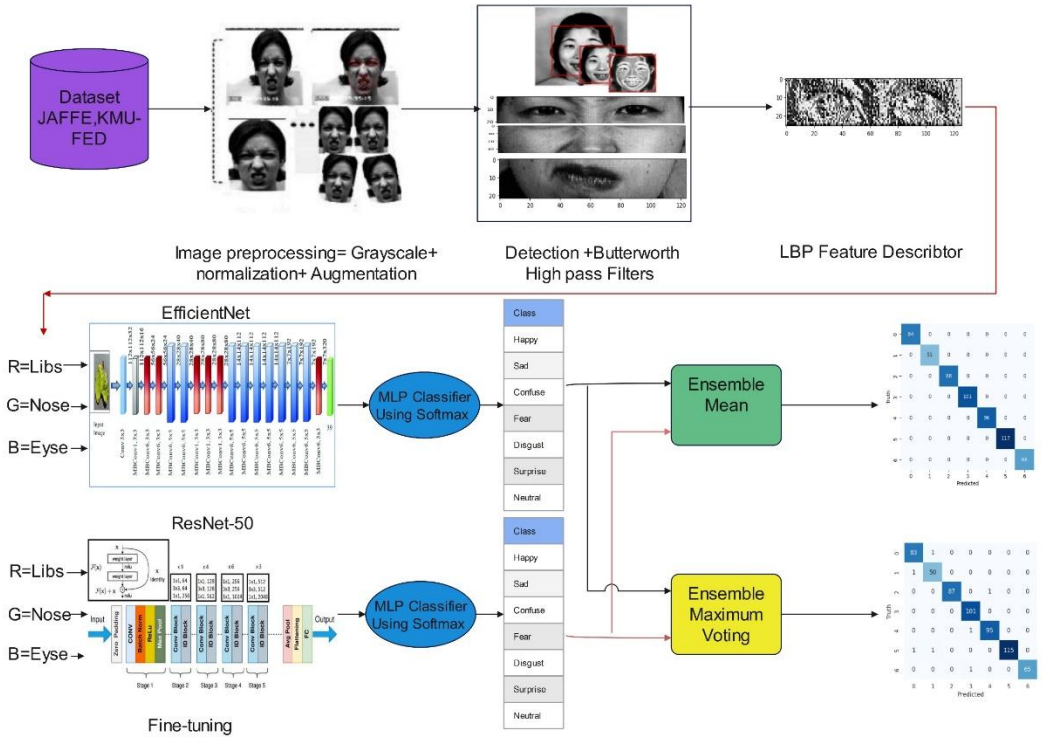

Fig. 1 Overall system block diagram

In Figure 1, the structure of the proposed method is depicted, demonstrating that ResNet-50 and EfficientNet are employed in parallel. The outputs of these two networks determine the final result using the Ensemble Mean and Ensemble Maximum Voting techniques.

3.2 Face detection

The Viola-Jones algorithm was used for face detection on grayscale images. As illustrated in Figure 2, the stages of the Viola-Jones algorithm were identified as the region of interest. The cropped rectangular shape varies in clarity. The spatial normalization process has changed the size of the cropped area to 256 x 256 pixels as a result. The spatial normalization process facilitates the operation of the FER system. Through the use of a high-pass Butterworth filter, the intensity values of the identified face region are enhanced. The areas of the eyes, nose, and mouth are detected using the Viola-Jones algorithm [26].

3.3 LBP feature descriptor

The Local Binary Pattern (LBP) approach has been utilized in various applications, including human face detection and facial expression recognition using the LBP algorithm. The LBP histogram is derived from the Gabor map of the human Face. A unified vector is then

Fig. 2 Four Stages of Viola-Jones Algorithm [26]

formed from the combined histograms. The vector is then referred to as a pattern vector. The LBP feature descriptor is widely used as a robust brightness-invariant feature descriptor. In this process, the operator compares the values of neighboring pixels with the values of the central pixel to determine a binary number. For 3×3 neighborhoods, the LBP operator is defined for each pixel acting as the central pixel and eight surrounding pixels are evaluated according to specified thresholds. As a result of the local neighboring matrix bits associated with each pixel, eight pixels surrounding each pixel are created. LBP descriptors are illustrated in Figure 3, where 8 bits are combined to form a binary number that is then converted into a decimal number. For each pixel (p), the 8 neighboring central pixels are compared with pixel (p), and if x is more significant than pixel (p), then the neighbors of the pixel receive a value of 1. It is possible to obtain a binary number by concatenating all these binary codes clockwise, starting at the top left [27].

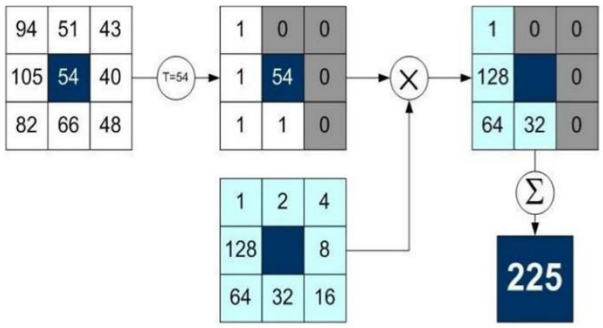

Fig. 3 An example of the basic LBP operator [27]

3.4 Convolutional neural networks

An artificial neural network that extracts features and processes them by overlapping convolutional layers and downsampling layers is known as a Convolutional Neural Network (CNN). Convolutional Neural Networks (CNNs) are perceptron-based models that can automatically extract features from images. Convolutional Neural Networks have become a hot research topic, and Figure 4 illustrates its structure. A significant advantage of CNNs is that they can receive the original images directly without excessive preprocessing. The Convolutional Neural Network (CNN) reduces the complexity of models by utilizing both the local and global information of a picture; it is capable of strong translation, rotation, and scaling capabilities. Residual neural networks ResNet-50 and EfficientNet were the classic convolutional neural networks used in this study [28].

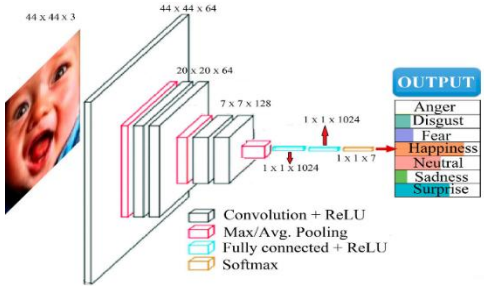

Fig. 4 The framework for custom CNN based Facial expression recognition [28].

3.4.1 ResNet

According to theoretical considerations, the accuracy of neural networks should increase by adding more layers. As a result of increasing network depth, the network accuracy tends to saturate and then rapidly deteriorate. It is commonly referred to as the vanishing gradient problem. Overfitting is not the sole cause of this problem. The root cause lies in deep neural networks' vanishing or exploding gradients. As a result of repeated multiplication during the backpropagation phase, gradients become infinitely small, resulting in minuscule parameter changes. Before the presentation of the remaining neural network, the intermediate normalization layers were initialized using a normalized initialization procedure. The Residual Neural Network (ResNet) is an architecture based on a CNN with a residual block as its main building block. To reduce vanishing gradients, residual blocks employ skip connections, which are connections that skip one or more layers. The residual shortcut ensures the integrity of the network if the regularization coefficient converges to zero in the training phase [29].

3.4.2 EfficienNet

The EfficientNet architecture is an innovative convolutional neural network architecture designed to enhance the efficiency and performance of CNNs. It combines three primary factors to achieve this.

Scaling

It improves the network's efficiency and performance by increasing its dimensions, depth, width, and resolution. The dimension of the network refers to the size of the image or input of the network. As the network's dimensions increase, the number of pixels in the network input will increase. The depth of the network is determined by the number of layers contained within it. The number of network parameters increases as the depth of the network increases. The width of the network represents the number of channels in each layer. The number of parameters in a network increases as the width of the network increases. The number of pixels in each channel signifies the resolution of the network. Scaling is an integral part of improving the efficiency and performance of convolutional neural networks by increasing the network resolution, which increases the number of network parameters. It is, however, important to select a combination of the network's dimensions, depth, width, and resolution to maximize its efficiency and performance.

Efficiency

Efficiencies in the EfficientNet architecture are achieved by employing techniques to enhance the network's efficiency without compromising performance. These techniques may include the following:

Utilizing thinner layers: Thinner layers process fewer data packets and can increase the efficiency of a network.

Compression of data: A network can process more data quickly and achieve greater efficiency by compressing data.

Parameter reduction: Reducing the number of parameters can reduce the computational workload on the network and improve its performance

Relative Efficiency

An EfficientNet score is a new metric that compares the efficiency of one network with that of other networks in the EfficientNet architecture.

$$\text{EfficientNet Score} = \frac{\text{Accuracy}}{TP+FN(\text{FLOPS} * \text{Parameters})} \quad (1)$$

FLOPS indicates the network's number of computations per second for this formula. A novel optimization function is employed in the EfficientNet architecture to locate a combination of network dimensions, depths, widths, and resolutions that maximize the relative efficiency of the network [30].

3.4.3 fine-tuning

In this method, we train all layers of each model, including the new output layers (Figure 5). The article selected two CNN models because they provided the highest accuracy in the evaluation phase [31].

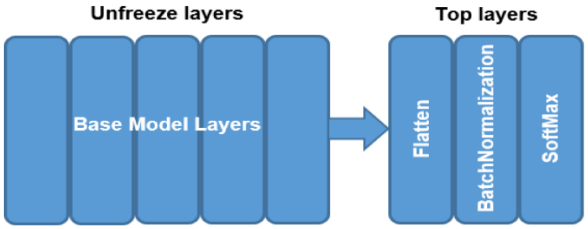

Fig. 5 Fine-tuning layout [31].

3.5 Ensemble learning

In ensemble learning, multiple base networks are observed, and their outputs are combined and integrated using rules. The rule used to combine the outputs determines the effectiveness of a group. This article uses two approaches to incorporate the outputs of learning models.

3.5.1 Ensemble Mean

The Ensemble mean of outputs of base networks in a composite model is an approach commonly used to merge or integrate output decisions. To obtain the final prediction of the ensemble model, the results of different neural networks are averaged. As deep learning architectures are characterized by high variance and low bias, simply averaging them enhances generalization performance by reducing variance across models. The average is computed either directly from the outputs of the base networks or based on the predicted class probabilities using the Softmax function [32].

$$P_i^j = softmax^j(O_i) = \frac{o_i^j}{\sum_{k=1}^{k} exp\,(o_k^j)} \quad (2)$$

There are three components to the prediction model: P_i^j the predicted probability of class (i) in the base learning network (j), O_i^j The output of class (i) in the base learning network (j), and (k) the number of classes in the base learning network. If a base learning network's performance and efficiency are comparable, Mean is a reasonable choice. Each deep learning network must perform well and efficiently to achieve the best results in combination with these networks. Thus, weak deep-learning networks cannot produce better results in ensemble learning [33].

3.5.2 Ensemble Maximum Voting

In the method of maximum voting, test samples are assigned to a class based on the outputs of various essential deep-learning networks. The majority vote measures the class output of each deep learning network as an alternative to considering average probabilities. It predicts the final label as the majority instead of considering average probabilities. Using this method, each classifier's opinion regarding the class of the input pattern is viewed as a vote, and the final decision is made based on the total number of votes gathered from the different classifiers. When the classifiers are independent, and their classification accuracy exceeds fifty percent, regardless of the number of classes, increasing the number of classifiers increases the accuracy of the voting method [34].

4 Result

4.1 Dataset

The JAFFE database pertains to a database of Japanese women. It is a grayscale dataset collected from psychological experiments. This dataset comprises 213 images of seven different (FE) collected in a controlled environment by a laboratory. It represents a combination of various facial expressions. Figure 6 illustrates seven images from the dataset [13].

Fig. 6 JAFFE Dataset [13].

The introduced dataset is not limited to a specific group of individuals; therefore, the proposed method has been evaluated using the KMU-FED dataset, which

contains images of drivers. The drivers' facial states were collected using a camera mounted on the dashboard or on the vehicle's steering wheel. This dataset's total number of images is 1106, categorized into six classes per the classification system introduced for basic emotions. These images display characteristics such as light reflections from different directions and obstacles such as hair and glasses in front of the Face. Figure 7 illustrates several examples [12].

Fig. 7 KMU-FED Dataset [12].

4.2 Performance Parameters

The evaluation and analysis of a model is a crucial step following its construction and design. In the following steps, the results are categorized into two groups: positive and negative. Using relevant indicators, the algorithm can then be assessed for quality. In terms of categorization, the data can be divided into four groups after analysis. False positives include (false positives, negatives that are classified as positives), true negatives (negatives that are classified as negatives), true positives (positive and classified as positive), and false negatives (positives that are classified as negatives). To evaluate various performance parameters, (3) to (6) provide expressions.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{5}$$

$$F1_{\_Score} = \frac{2*\text{Precision}* Recall}{\text{Precision}+ Recall} \tag{6}$$

4.3 Experiment on the Dataset

Several studies have been conducted on images from the JAFFE and KMU-FED databases in this paper. We have implemented group-based deep learning for facial expression recognition using the TensorFlow framework and the Keras library on the Google Colab service. The libraries are first imported, followed by preprocessing steps, which normalize and convert images to grayscale. Additionally, data augmentation techniques have been used to prevent overfitting. As a result, each image's rotations and orientations must be created to provide sufficient data for the training phase. In addition, the use of Butterworth filters enhances the quality of the image to ensure a higher level of accuracy during the facial recognition phase. Figure 8 illustrates a Butterworth filter by detecting and cropping images using the Viola-Jones face detection algorithm. Using facial landmarks, classes are written using multiple methods utilizing the Viola-Jones face detection algorithm. The classes are invoked using the initialize method, and a library for face mesh is imported, which identifies the facial landmarks. It is then necessary to identify the landmarks for lips, noses, and eyes and extract them using the get method, as shown in Figure 9. These steps lead to generating Local Binary Patterns (LBP) from each image, which are then fed into one of three deep neural network channels for each component (eyes, mouth, nose). As shown in Figure 10, the final preprocessed output of the proposed method utilizes two convolutional neural networks, ResNet-50 and EfficientNet. Figure 10 presents the final preprocessed output, where the lips, nose, and eyes have been extracted sequentially from the RGB channels of the ResNet-50 and EfficientNet networks. GlobalAveragePooling is used to construct the two investigated networks. The convolutional layer's output, a multidimensional tensor, is transformed into a one-dimensional tensor by incorporating a Dense flattened layer with 512 neurons, forming a fully connected network. The second dense layer consists of 256 neurons connected to the previous layer's output. This results in creating a thick layer with 7 neurons, corresponding to the number of classes. For data classification, a set of probabilities is required for final decision-making, though a deep neural network uses many layers to comprehend different aspects of the data. Softmax is a well-known function that normalizes probability values within a standard range of 0 to 1.

The dropout technique is utilized to prevent overfitting. It is applied between the first and second layers to apply a dropout of 0.3, which excludes 30% of the neurons. "The first and second layers" refer to the layers that follow the preprocessing stage. A shrinking network ultimately remains a result of each neuron being removed from the network at each training stage with a probability of 1-P or retained with a probability of P at each training stage. As a result, input and output connections to a node are also eliminated, ensuring that only the reduced network can be trained using the data at that point. Finally, the weights are optimized using the Adam optimizer, widely used in TensorFlow for computing various optimization functions.

Each neural network assigns input patterns to a class as part of the maximum voting process. Based on the majority vote among the two neural networks, the input pattern will be transferred to the final class based on the

majority vote of the two neural networks. With the averaging rule, each neural network will assign a probability to seven classes based on the output of the softmax function. As a result of averaging the outputs of the softmax functions from the two networks, the class whose outputs exceed the highest value will receive the input pattern. In the results section, it will be observed that the averaging method outperforms the voting rule in performance.

The hyperparameters of the ResNet-50 and EfficientNet neural networks used are as follows:

Table 1 Model Hyperparameters

Hyperparameter	Description
Learning Rate	5e-2
Batch Size	8
Number of Epochs	100
Optimizer	Adam
Dropout Rate	0.3
Activation Function	relu

4.3.1 Experiments on the JAFFE dataset

Following data augmentation, 1243 images were used, with a 20 to 80 percent ratio for training and test data in this stage. In particular, 970 images were assigned to training, whereas 243 were allocated to testing. 94.65 percent and 97.94 percent accuracy of facial expression recognition were achieved with two deep networks, RessNet-50 and EfficientNet, respectively. Following the fusion of the deep learning networks, two different fusion rules were applied, namely Ensemble Mean and Ensemble Maximum Voting. An ensemble maximum voting achieved a facial expression recognition accuracy of 97 percent, while an ensemble Mean achieved a 98 percent accuracy. Figure 11 illustrates the confusion matrix corresponding to the proposed structures. Table 1 presents the performance parameters for ResNet-50, EfficientNet, Ensemble Maximum Voting, and Ensemble Mean.

4.3.2 Facial Expression Recognition Results on the JAFFE Dataset using Confusion Matrix

As a tool for evaluating classification systems, the confusion matrix is one of the most effective tools. Each row and column corresponds to a class of data. Each row indicates the number of input data that the system has assigned to each class. If all data are correctly classified, the confusion matrix will be diagonal, and elements other than the main diagonal will be zero. The closer the matrix is to a diagonal matrix, the higher the system's accuracy. The confusion matrix is calculated to evaluate the performance of the proposed method.

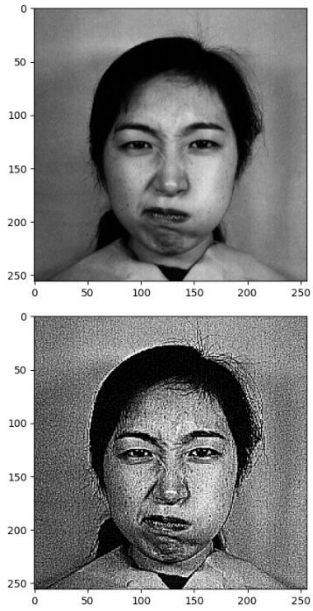

Fig. 8 Butherworth highpass filter

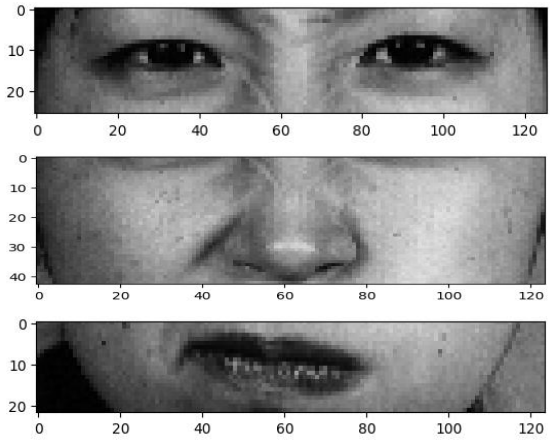

Fig. 9 Sample detected eye, nose, mouth regions

Table 2 Comparison of Performance Parameters (JAFFE)

Model	Accuracy	Precision	Recall	F1-Score
ResNet-50	94.65	94.86	94.14	94.14
EfficienNet	97.94	97.86	97.43	97.57
Ensemble voting	97	96.85	97.14	96.86
Ensemble mean	99	99	98.86	99

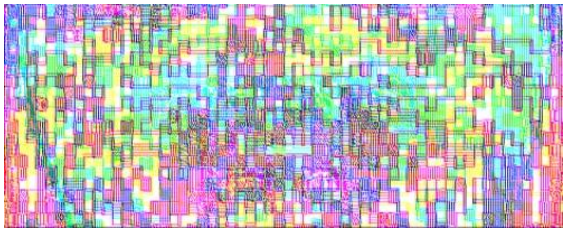

Fig. 10 Preprocessing Conducted on Images

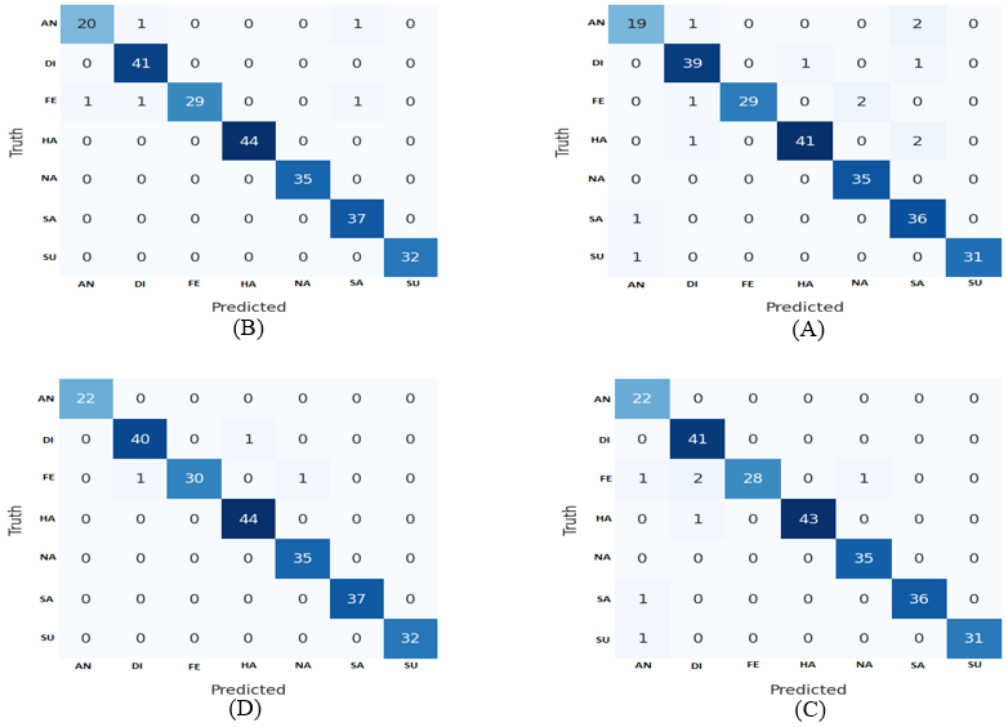

Fig. 11 Confusion Matrix (JAFFE) A) RessNet-50 B) EfficientNet C) Ensemble Maximum Voting D) Ensemble Mean

The confusion matrix resulting from the classification of the JAFFE dataset is shown in Figure 11. As a result of more distinctive features in the Face, the recognition of happiness is usually easier than other emotional states in the recognition of different emotions. As Sadness is more easily confused, Izard believes it is similar to other emotions, such as anger and fear. There is a tendency for the facial expressions of anger, fear, and Sadness to appear identical, resulting in potential misclassifications. As observed, the recognition of anger and fear states has been accompanied by more errors than other classes.

4.3.3 Performance Comparisons with the Previously Reported Techniques on the JAFFE Dataset

As shown in Table 2, various techniques on the JAFFE database were utilized for comparative results. Sajjanhar et al. [35] used publicly available facial databases, including CK+, JAFFE, and FACES, to train a deep CNN model for facial expression recognition. Based on pre-trained models, Inception V3, VGG19, and VGG-Face, accuracies were reported to be 75.88%, 94.71 %, and 86.67%, respectively. Similarly, Kartheek et al. [36] reported FER accuracy of 66.2% on the JAFFE dataset using the SVM technique with feature descriptors based on Windmill Gaussian Finite Difference (WGFD) descriptors. With Deep Subspace Feature Learning (DSFL) based on PCANet and LDANet models, Sun et al. [37] achieved accuracy rates of 71.38 percent and 70.18 percent, respectively. According to Sahoo et al., pre-trained models AlexNet, SqueezeNet, and VGG19 achieved accuracy rates of 66.52, 57.8%, and 84.4%, respectively. In contrast, Bhatti et al. [38] and Minei et al. [1] show a performance of more than 91% accuracy.

Table 3 Summary of results on JAFFE

Reference	Model	Accuracy
Sajjanhar et al. [35].	Pre-trained InceptionV3	75.88
	Pre-trained VGG19	94.71
	Pre-trained VGG-Face	86.67
Bhatti et al.[38]	RELM	91.67
Minaee et al. [1]	Attentional CNN	92.8
Kartheek et al.[36]	SVM	66.2
Sun et al. [37]	PCANet	71.38
	LDANet	70.18
Goutam Kumar Sahoo et al.[11]	Pre-trained AlexNet	66.52
	Pre-trained SqueezeNet	57.8
	Pre-trained VGG19	84.4
Ghasemi et al.	RessNet-50	94.65
	EfficientNet	97.94
	Ensemble Maximum Voting	97
	Ensemble Mean	99

The performance of the proposed work demonstrates better results compared to other competitors. In evaluating the performance of the proposed method, the accuracies of the models RessNet-50, EfficientNet, Ensemble Maximum Voting, and Ensemble Mean were 94.65%, 97.94%, 97%, and 99%, respectively.

4.3.4 Experiments on the KMU-FED dataset

The KMU-FED database, 1101 images were allocated for training and 221 for testing with an 80 to 20 train-to-test ratio. The two deep networks, ResNet-50 and EfficientNet, were initially used to achieve 97% and 92% facial expression recognition accuracy, respectively. A combination of deep learning networks and ensemble methods, such as Ensemble Maximum Voting and Ensemble Mean of classifiers, was then used to achieve facial expression recognition accuracy of 95% and 98%, respectively, with Ensemble Maximum Voting and Ensemble Mean. Figure 12 illustrates the confusion matrix describing the proposed structures. Table 3 presents the evaluation metrics for ResNet-50, EfficientNet, Ensemble Maximum Voting, and Ensemble Mean.

Table 4 Comparison of Performance Parameters (KMU-FED)

Model	Accuracy	Precision	Recall	F1-Score
ResNet-50	97	97	96.66	96.66
EfficienNet	92	92.16	90.83	91.33
Ensemble Maximum Voting	95	94.83	95.16	94.83
Ensemble Mean	98	98.50	98.50	98.50

4.3.5 Facial Expression Recognition Results on the KMU-FED Dataset using Confusion Matrix

The evaluation of the proposed structure of the KMU-FED database is illustrated in Figure 12. In this database, the proposed structure has encountered errors due to the similarity between the emotions of fear and anger in some cases. In addition, in the EfficientNet and Ensemble Maximum Voting structures, the error rate for the happy emotion is higher than in other structures. Overall, the proposed structure has performed well on this database.

Table 5 Summary of results on KMU-FED

Reference	Model	Accuracy
Jeong et al.[6]	LMRF	95.1
	MobileNetV3	94.9
	SqueezeNet	89.7
Jeong and Chul Ko[11]	WRF	94.7
Leone et al.[39]	VGG16	94.2
J. Zhang et al. [40]	CCNN	97.3
Ghasemi et al.	RessNet-50	97
	EfficientNet	92
	Ensemble Maximum Voting	95
	Ensemble Mean	98

4.3.6 Performance Comparisons with the Previously Reported Techniques on the KMU-FED Dataset

The KMU-FED database consists explicitly of images of drivers, and since this research aims to enhance drivers' safety, evaluating the proposed method on this database would be helpful. The technology enables the implementation of real-time applications with embedded devices in vehicles by leveraging machine learning and deep learning. Based on the evaluation results of the proposed method on this database, Table 4 compares it with other deep learning- or machine learning-based methods, such as MobileNetV3 [6], VGG [39], SqueezeNet [6], CCNN [40], LMRF [6], and WRF [11] in terms of recognition accuracy.

In Figure 13, a set of experimental samples is accompanied by their true or correct labels and the labels that the model attributes or predicts for the input samples. The "actual label" refers to the true label, while the "predicted label" represents the label identified by our model as belonging to a specific class of facial expressions.

As part of my article, I would like to examine the significance and applications of the JAFFE and KMU-FED datasets. These two datasets, with their high diversity of facial expressions, have enabled me to analyze driver emotions in advanced driver-assistance systems thoroughly.

The JAFFE dataset comprises images of faces exhibiting various emotional states, facilitating the recognition and analysis of emotions. Conversely, the KMU-FED dataset provides additional information regarding facial behavior in driving situations. By integrating these two data sources, I have achieved findings that not only encompass a wide range of emotional expressions but also contribute to enhancing human-machine interactions.

Research indicates that understanding driver emotions can empower driver-assistance systems to make more intelligent decisions in specific scenarios. For instance, if the system detects a driver experiencing stress or fatigue, it could implement measures such as issuing alerts or automatically adjusting speed.

Ultimately, this research could serve as a foundation for future investigations and assist in developing automated technologies focusing on human and emotional interactions. I hope that the results of this work lead to further advancements in the field of driver-assistance systems.

5 Conclusions

This article proposes a deep learning framework that utilizes Local Binary Patterns (LBP) feature extraction for driver assistance applications in vehicles can be highly effective. Therefore, the results of the proposed

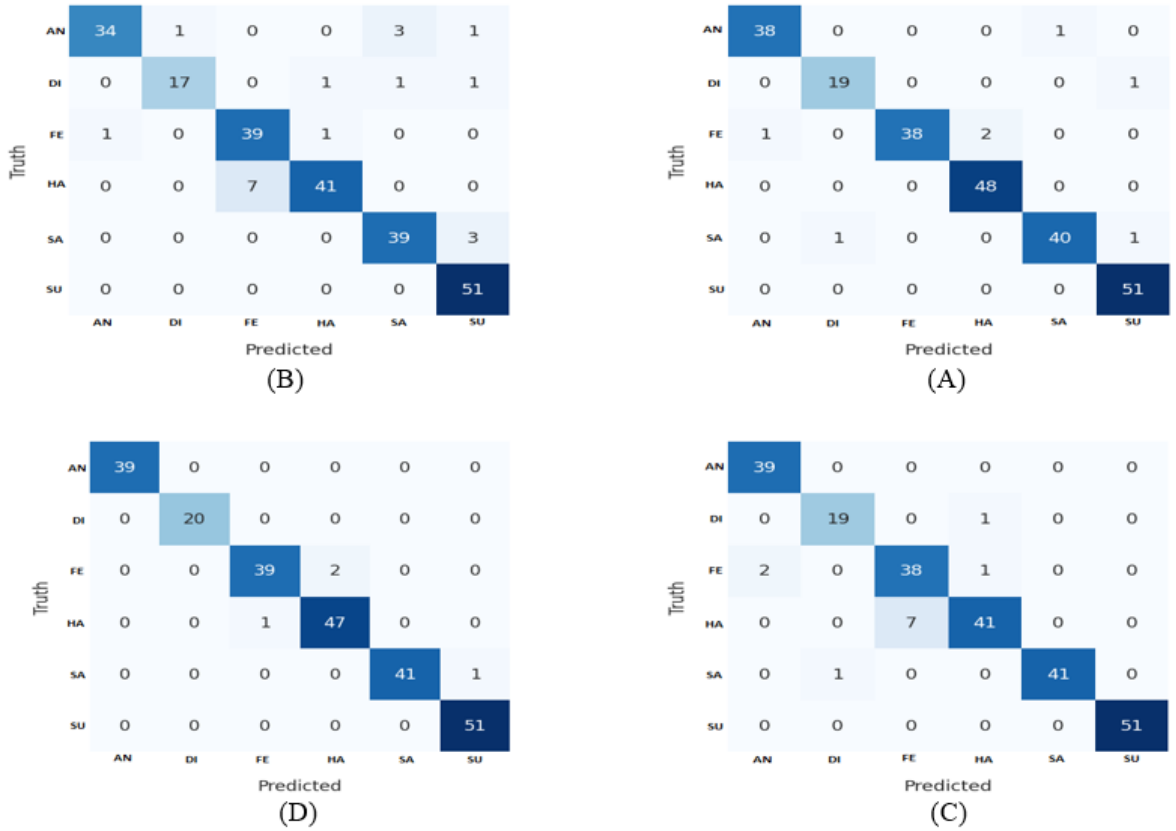

Fig. 12 Confusion Matrix (KMU-FED) A) RessNet-50 B) EfficientNet C) Ensemble Maximum Voting D) Ensemble Mean

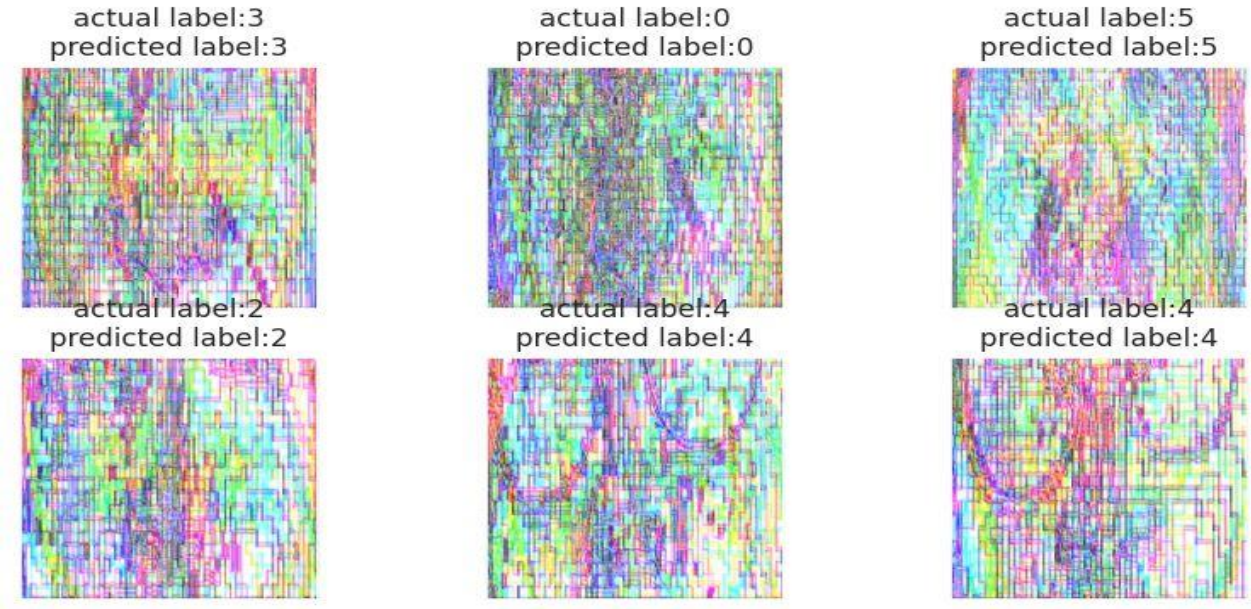

Fig. 13 Experimental validation of the proposed model

LBP algorithm can be used to identify human emotion in various facial regions, considering facial expressions. The ensemble of classifiers enhances overall classification performance. Ensemble Maximum Voting and Ensemble Mean have been utilized to make final decisions about experimental samples within their

respective classes. As shown in Tables 1 and 2, the proposed method addresses the shortcomings of classifiers, and utilizing the committee rule can significantly improve accuracy. In this study, it is noted that combining traditional methods in the preprocessing stage and using ensemble classifier techniques in deep neural networks has resulted in the proposed method achieving higher accuracy compared to its competitors.

Furthermore, selecting deep neural networks has necessitated various research and experiments to identify the most suitable deep learning network for this task. Consequently, adherence to all these factors has led to the development of a new model with high accuracy, which can be applied in sensitive and high-demand areas such as autonomous driving. In the future, autoencoders may be used to reduce the features of deep neural networks using Graphics Processing Units (GPUs) to enhance the speed and accuracy of the proposed algorithm.

References

[1] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 3046, 2021, doi: 10.3390/s21093046.

[2] E. Ghasemi Bideskan, S. M. Razavi, S. Mohamadzadeh, and M. Taghippour, "Facial Expression Recognition through Optimal Filter Design Using a Metaheuristic Kidney Algorithm," *J. Electr. Comput. Eng. Innov.*, vol. 12, no. 2, pp. 425–438, 2024.

[3] S. Mohamadzadeh, S. Pasban, J. Zeraatkar-Moghadam, and A. K. Shafiei, "Parkinson's disease detection by using feature selection and sparse representation," *J. Med. Biol. Eng.*, vol. 41, no. 4, pp. 412–421, 2021.

[4] C. Nandagopal, P. Anisha, K. G. Dharani, and N. Kuraloviya, "Smart Accident Detection and Rescue System using VANET," in *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, 2022, pp. 1111–1116.

[5] M. Rohani, H. Farsi, and S. Mohamadzadeh, "Deep Multi-task Convolutional Neural Networks for Efficient Classification of Face Attributes," *Int. J. Eng.*, vol. 36, no. 11, pp. 2102–2111, 2023.

[6] M. Jeong, J. Nam, and B. C. Ko, "Lightweight multilayer random forests for monitoring driver emotional status," *Ieee Access*, vol. 8, pp. 60344–60354, 2020.

[7] S. Li and W. Deng, "Deep facial expression recognition: a survey," *Journal of Image and Graphics*, vol. 25, no. 11. abs, pp. 2306–2320, 2020, doi: 10.11834/jig.200233.

[8] Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial expression recognition: A survey," *Symmetry (Basel).*, vol. 11, no. 10, p. 1189, 2019.

[9] R. Elhabob, Y. Zhao, I. Sella, and H. Xiong, "An efficient certificateless public key cryptography with authorized equality test in IIoT," *J. Ambient Intell. Humaniz. Comput.*, vol. 11, pp. 1065–1083, 2020.

[10] N. Kumar and D. Dash, "Flow based efficient data gathering in wireless sensor network using path-constrained mobile sink," *J. Ambient Intell. Humaniz. Comput.*, vol. 11, no. 3, pp. 1163–1175, 2020.

[11] G. K. Sahoo, S. K. Das, and P. Singh, "Performance Comparison of Facial Emotion Recognition: A Transfer Learning-Based Driver Assistance Framework for In-Vehicle Applications," *Circuits, Syst. Signal Process.*, pp. 1–28, 2023.

[12] R. Singh, S. Saurav, T. Kumar, R. Saini, A. Vohra, and S. Singh, "Facial expression recognition in videos using hybrid CNN & ConvLSTM," *Int. J. Inf. Technol.*, vol. 15, no. 4, pp. 1819–1830, 2023.

[13] Y. DAŞDEMİR and R. Özakar, "Affective states classification performance of audio-visual stimuli from EEG signals with multiple-instance learning," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 30, no. 7, pp. 2707–2724, 2022.

[14] C. Karras, A. Karras, and S. Sioutas, "Pattern recognition and event detection on IoT data-streams," *arXiv Prepr. arXiv2203.01114*, 2022.

[15] A. Akbari, H. Farsi, and S. Mohamadzadeh, "Deep neural network with extracted features for social group detection," *J. Electr. Comput. Eng. Innov.*, vol. 9, no. 1, pp. 47–56, 2021.

[16] P. D. Marrero Fernandez, F. A. Guerrero Pena, T. Ren, and A. Cunha, "Feratt: Facial expression recognition with attention net," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, p. 0.

[17] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4057–4069, 2020.

[18] A. Gupta, S. Arunachalam, and R. Balakrishnan, "Deep self-attention network for facial emotion recognition," *Procedia Comput. Sci.*, vol. 171, pp. 1527–1534, 2020.

[19] Y. Gan, J. Chen, Z. Yang, and L. Xu, "Multiple attention network for facial expression recognition," *IEEE Access*, vol. 8, pp. 7383–7393, 2020.

[20] C. Wang, J. Ding, H. Yan, and S. Shen, "A prototype-oriented contrastive adaption network for cross-domain facial expression recognition," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 4194–4210.

[21] M. KHALID, M. KEMING, and T. HUSSAIN, "Design and implementation of clothing fashion style recommendation system using deep learning," *Rev. Română Informatică și Autom.*, vol. 31, no. 4, 2023, doi: 10.33436/v31i4y202110.

[22] M. I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64827–64836, 2019, doi: 10.1109/ACCESS.2019.2917266.

[23] D. Kollias and S. Zafeiriou, "Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface," *arXiv Prepr. arXiv1910.04855*, 2019.

[24] Irfanullah, T. Hussain, A. Iqbal, B. Yang, and A. Hussain, "Real time violence detection in surveillance videos using Convolutional Neural Networks," *Multimed. Tools Appl.*, vol. 81, no. 26, pp. 38151–38173, 2022.

[25] K. Zaman, Z. Sun, S. M. Shah, M. Shoaib, L. Pei, and A. Hussain, "Driver emotions recognition based on improved faster R-CNN and neural architectural search network," *Symmetry (Basel).*, vol. 14, no. 4, p. 687, 2022.

[26] D. Lakshmi and R. Ponnusamy, "Facial emotion recognition using modified HOG and LBP features with deep stacked autoencoders," *Microprocess. Microsyst.*, vol. 82, p. 103834, 2021.

[27] Y. ELsayed, A. ELSayed, and M. A. Abdou, "An automatic improved facial expression recognition for masked faces," *Neural Comput. Appl.*, vol. 35, no. 20, pp. 14963–14972, 2023.

[28] R. A. Borgalli and S. Surve, "Review on learning framework for facial expression recognition," *Imaging Sci. J.*, vol. 70, no. 7, pp. 483–521, 2022.

[29] Y. Djenouri, A. Belhadi, A. Yazidi, G. Srivastava, and J. C. Lin, "Artificial intelligence of medical things for disease detection using ensemble deep learning and attention mechanism," *Expert Syst.*, p. e13093, 2022.

[30] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019, pp. 6105–6114.

[31] N. S. Abdulsattar and M. N. Hussain, "Facial expression recognition using transfer learning and fine-tuning strategies: A comparative study," in *2022 International Conference on Computer Science and Software Engineering (CSASE)*, 2022, pp. 101–106.

[32] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *Eng. Appl. Artif. Intell.*, vol. 115, p. 105151, 2022.

[33] T. Zhou, H. Lu, Z. Yang, S. Qiu, B. Huo, and Y. Dong, "The ensemble deep learning model for novel COVID-19 on CT images," *Appl. Soft Comput.*, vol. 98, p. 106885, 2021.

[34] D. Müller, I. Soto-Rey, and F. Kramer, "An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks," *Ieee Access*, vol. 10, pp. 66467–66480, 2022.

[35] M. Khalid, J. Baber, M. K. Kasi, M. Bakhtyar, V. Devi, and N. Sheikh, "Empirical evaluation of activation functions in deep convolution neural network for facial expression recognition," in *2020 43rd International conference on telecommunications and signal processing (TSP)*, 2020, pp. 204–207.

[36] M. N. Kartheek, M. V. N. K. Prasad, and R. Bhukya, "Windmill graph based feature descriptors for facial expression recognition," *Optik (Stuttg).*, vol. 260, p. 169053, 2022.

[37] Z. Sun, H. Zhang, S. Ma, and Z. Hu, "Combining filtered dictionary representation based deep subspace filter learning with a discriminative classification criterion for facial expression recognition," *Artif. Intell. Rev.*, vol. 55, no. 8, pp. 6547–6566, 2022.

[38] Y. K. Bhatti, A. Jamil, N. Nida, M. H. Yousaf, S. Viriri, and S. A. Velastin, "Facial Expression Recognition of Instructor Using Deep Features and Extreme Learning Machine," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–17, 2021, doi: 10.1155/2021/5570870.

[39] A. Leone, A. Caroppo, A. Manni, and P. Siciliano, "Vision-based road rage detection framework in automotive safety applications," *Sensors*, vol. 21, no. 9, p. 2942, 2021.

[40] J. Zhang, X. Mei, H. Liu, S. Yuan, and T. Qian, "Detecting negative emotional stress based on facial expression in real time," in *2019 IEEE 4th international conference on signal and image processing (ICSIP)*, 2019, pp. 430–434.

Ehsan Ghasemi received his AD.Sc. degree in Electrical Engineering from university of islamic republic of iran broadcasting, in 2014 and his B.Sc. degree in Electrical Engineering from Islamic Azad University, Birjand, in 2018 and M.Sc. degree in Electrical Engineering from the Birjand University, in 2021 respectively. He is currently a Ph.D. student at Birjand University to receive a Ph.D. degree in electronics engineering. His research interests include Computer Vision, Pattern Recognition, optimization algorithms and Artificial Intelligence.

Seyyed Mohammad Razavi received his B.Sc. degree in Electrical Engineering from Amirkabir University of Technology, in 1994 and his M.Sc. and Ph.D. degree in Electrical Engineering from the Tarbiat Modares University, Iran, in 1996 and 2006 respectively. Now, he is a Full Professor in University of Birjand. His research interests include Computer Vision, Pattern Recognition and Artificial Intelligence.

Sajad Mohamadzadeh received the B.Sc. degree in communication engineering from Sistan & Baloochestan, University of Zahedan, Iran, in 2010. He received the M.Sc. and Ph.D. degree in communication engineering from South of Khorasan, University of Birjand, Birjand, Iran, in 2012 and 2016, respectively. Now, he works as associte professor at department of electrical and computer engineering, University of Birjand, Birjand, Iran. His area research interests include Image and Video Processing, Deep Neural Network, Pattern recognition, Digital Signal Processing, Sparse Representation, and Deep Learning.