Complex & Intelligent Systems
https://doi.org/10.1007/s40747-023-01100-9

ORIGINAL ARTICLE

A novel driver emotion recognition system based on deep ensemble classification

Khalid Zaman[1] · Sun Zhaoyun[1] · Babar Shah[2] · Tariq Hussain[3] · Sayyed Mudassar Shah[1] · Farman Ali[4] · Umer Sadiq Khan[5,6]

Received: 5 November 2022 / Accepted: 1 May 2023

Abstract
Driver emotion classification is an important topic that can raise awareness of driving habits because many drivers are overconfident and unaware of their bad driving habits. Drivers will acquire insight into their poor driving behaviors and be better able to avoid future accidents if their behavior is automatically identified. In this paper, we use different models such as convolutional neural networks, recurrent neural networks, and multi-layer perceptron classification models to construct an ensemble convolutional neural network-based enhanced driver facial expression recognition model. First, the faces of the drivers are discovered using the faster region-based convolutional neural network (R-CNN) model, which can recognize faces in real-time and offline video reliably and effectively. The feature-fusing technique is utilized to integrate the features extracted from three CNN models, and the fused features are then used to train the suggested ensemble classification model. To increase the accuracy and efficiency of face detection, a new convolutional neural network block (InceptionV3) replaces the improved Faster R-CNN feature-learning block. To evaluate the proposed face detection and driver facial expression recognition (DFER) datasets, we achieved an accuracy of 98.01%, 99.53%, 99.27%, 96.81%, and 99.90% on the JAFFE, CK+, FER-2013, AffectNet, and custom-developed datasets, respectively. The custom-developed dataset has been recorded as the best among all under the simulation environment.

Keywords Driver facial expression recognition (DFER) · Custom developed datasets (CDD) · Computer vision · Attention mechanism and DenseNet · FE

Introduction

In computer vision and artificial intelligence, the facial expression is a significant and promising field of research and one of the primary processing methods for intentions expressed through nonverbal means. When interacting with

Khalid Zaman and Farman Ali have contributed equally to this work and are the first co-authors.

✉ Sun Zhaoyun
chysun@chd.edu.cn

✉ Tariq Hussain
uom.tariq@gmail.com

Khalid Zaman
khalidzaman@chd.edu.cn

Babar Shah
babar.shah@zu.ac.ae

Sayyed Mudassar Shah
mudassarshah@chd.edu.cn

Farman Ali
farmankanju@sejong.ac.kr

Umer Sadiq Khan
umersadiq@hbeu.edu.cn

[1] Information Engineering School, Chang'an University, Xi'an 710061, China

[2] College of Technological Innovation, Zayed University, Dubai 19282, UAE

[3] School of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou 310018, China

[4] Department of Computer Science and Engineering, School of Convergence, College of Computing and Informatics, Sungkyunkwan University, Seoul 03063, South Korea

[5] School of Computer and Information Science, Hubei Engineering University, Xiaogan Hubei 432000, China

[6] Institute for AI Industrial Technology Research, Hubei Engineering University, Xiaogan Hubei 432000, China

Published online: 07 June 2023

other people, it is impossible to avoid experiencing emotions [1]. They may or may not be visible to the naked eye. Therefore, trained professionals can detect and recognize any indication [2] before or after it is expressed if they have the appropriate tools at their disposal, regardless of when it occurs [3]. R-CNN and deep learning classifier techniques are used for emotion recognition. Only a few of the topics covered in this study are medical [4, 5], human–machine interfaces [6], urban sound perception [7], and animation [8]. Several fields, including the diagnosis of autism spectrum disorder in children [9] and security [10, 11], are seeing an increase in emotion recognition technology. To recognize emotions, various features such as EEG [9], facial expressions (FE) [5, 12, 13], text [14], and speech [15, 16] are used.

Moreover, due to various factors, including their ease of recognition, FE features are one of the most well-known methods of human recognition. The following are some advantages: (1) they are noticeable and visible; (2) they allow for the quick and easy collection of large face datasets using facial expression features; and (3) they contain a large number of features for emotion recognition [17, 18]. Through deep learning, specifically CNN-based learnable image features [15], it is also possible to compute, learn, and extract good facial expressions [19, 20]. Experts predict that FE will become increasingly significant due to advancements in artificial intelligence technology and the rising demand for applications in the era of big data. To be effective in complex environments, such as those characterized by occlusion, multiple views, and multiple objectives, facial emotion recognition (FER) solutions must be proposed in novel and innovative ways. The most relevant data collected under the most favorable conditions at the time of collection is highly desirable when attempting to accurately classify facial expressions to train a FE classifier [21]. The "golden rule" is a term used to describe this. Traditionally, a DFER system will first preprocess the image that will be used as an input to accomplish this. Face detection is a preprocessing step included in most peer-reviewed papers and is described in detail here. The nose and mouth are the most frequently used facial expression cues, even though numerous human face regions can be used to cue facial expressions. The cheeks, forehead, and eyes are just a few other parts of the face that can cue different types [22] of FE. According to a recent study, a small amount of data is collected from the ears and hairs when detecting FE [23].

Accordingly, since the mouth and eyes can detect more FE than the rest, the computer vision deep learning model should place the most significant emphasis on these parts of the face and ignore the others. A CNN framework for DFER is proposed in this manuscript [24], which is based on the findings of this study and incorporates some of the observations made above. Attentional mechanisms, in particular, are employed to [25] draw attention to the essential features of the face. Attentional convolutional networks can achieve extremely high accuracy rates even when using only a few layers (i.e., no more than 50).

Numerous techniques for extracting features have been utilized in the literature for recognizing emotions in drivers, but these techniques restrict the recognition and extraction of emotions. In addition, most emotion recognition systems rely on handcrafted features (such as grayscale statistics, RGB histogram, RGB statistical, and geometrical features) and traditional machine learning classifiers (e.g., SVM, K-NN, and Naive Bayes). The majority of work on feature extraction in recognition systems employs standard methods such as LPB, HOG, and GLCM. However, handcrafted features are considered less robust due to their non-invariant nature, and they extract too many features, which can negatively affect model training and validation. Some invariant features, such as SURF, SIFT, and ORB, operate on low-level features such as edges and corners. Nevertheless, due to the vast number of images in a dataset, these descriptors may not achieve high accuracy. Image data have various forms of noise, blurriness, oversharpening, and unbalanced contrast, which may impair model training, resulting in low accuracy.

This paper develops a deep learning model based on the improved Faster R-CNN and deep ensemble classifier for emotion recognition, along with the ability to observe the driver's emotions while driving a vehicle. The Faster R-CNN model has been improved for detecting the driver's face region, and the learning block of the Faster R-CNN has been replaced with an improved CNN block (InceptionV3), which improves the accuracy and efficiency of face detection. The proposed approach can work in environments where outdated and other approaches are deficient with full potential. The main contributions of this paper are as follows:

- A detailed systematic study is carried out on driver emotion recognition in videos and images.
- A custom driver emotion recognition dataset is developed, where the emotions of 30 drivers were recorded in challenging environments (illumination and obstacles).
- Face detection was performed using Improved Faster R-CNN.
- Transfer learning was performed using state-of-the-art updated CNN models such as DenseNet (201 layers).
- The proposed models were validated using various driver emotion benchmark datasets, including JAFFE, CK+, FER-2013, AffectNet, and the custom-developed dataset (CDD). To improve the accuracy of the model, data augmentation was used to expand the datasets.

The paper is organized as follows: the first section elaborates and explains the introduction of the entire manuscript. The next section explains and highlights the related work with

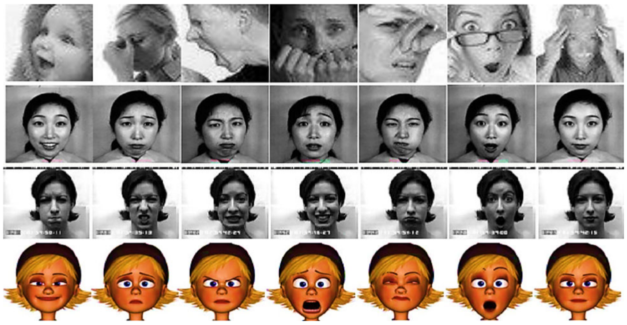


Fig. 1 Facial appearance and texture feature-based robust DFER framework for sentiment knowledge discovery

regard to current works from different authors' perspectives. The third section introduces the methodology and the proposed framework of the research. The fourth section explains and illustrates the experimental results of the overall research. Finally, the fifth section concludes the manuscript.

Related work

The six primary emotions: the emotions of pleasure, fear, hate, sorrow, disgust, and surprise (except neutral) are identified in [26]. Ekman utilized this concept to create *the facial action coding system* (FACS) [27], which became the gold standard for emotion recognition research. Neutral was later added to most human recognition datasets as a seventh fundamental emotion. Figure 1 shows sample pictures of various emotions from the four benchmark datasets (FER-2013, JAFFE, CK+, and AffectNet datasets). The primary emotions are the happy face, angry face, disgusted face, fearful face, sad face, surprised face, and contemptuous face.

A two-step machine learning methodology was employed in initial studies on emotion recognition. In the first step, the image's attributes are extracted, whereas, in the second step, classifiers are used for emotion detection. Gabor wavelets [28], Haar features [29], Texture features linear binary pattern (LBP) [30], and Edge Histogram Descriptors [31, 32] are some of the most frequently exploited manual features for the detection of FE. The classifier then identifies the image with the most appropriate sentiment. These techniques seem to be effective on more specialized datasets. However, posing significant limitations when applied to challenging datasets (with greater intra-class variation). To help the reader better understand some of the problems that images can provide, Fig. 1 of the first row included an image that only showed the reader's eyes or the covered hand or portions of the face.

Numerous firms have made significant advancements in neural networks, deep learning, picture categorization, and vision challenges. In [33], Khorrami demonstrated that CNN could achieve a better accuracy level for emotion recognition. Moreover, zero-bias CNN's Toronto Face Datasets (TFD) and Cohn–Kanade dataset (CK+) for attaining state-of-the-art results when applied to model human facial expressions. To construct a model for FE of stylized animation characters, the authors in [34] trained a network using deep learning and translated human images to animated faces. A FER neural network A network with a top layer of pooling, two convolution layers, and four initial layers or subnetworks has been proposed by Mollahosseini [35]. The authors in [36] integrate the removal and classification of features using a single recurrent network, highlighting the importance of input from both components. To achieve cutting-edge CK+ and JAFFE accuracy, the BDBN Network was utilized.

The authors in [37] implemented a deep CNN on noisy labeling of authentic images acquired through crowdsourcing. They deployed ten taggers to re-enact each image to acquire the required precision, with ten tags in their dataset and numerous costing functions for their Deep Convolutional neural network (DCNN). To improve the spontaneous recognition of facial experiences, authors in [38] used more discriminative neurons that outperformed *Incremental Boosting CNN* (IB-CNN). The authors of [39] An identity-aware CNN (IA-CNN) created that uses identity- and expression-sensitive contrast loss to reduce variation of expression-related information during identity learning. Similarly, they have developed a network architecture with a focal model called end-to-end network architecture [40]. To minimize uncertainty and avoid unclear face images (caused by labeling noise) from overfitting the deeper network, the authors in [41] devised a quick and efficient self-repair technique (SCN). SCN reduces uncertainty in two dimensions and ways: (1) using a self-attention mechanism to weight each sample of workouts in small batches with rank regularization; and (2) by carefully changing these samples in the lowest rank set. It was identified using an algorithm. In the real world, it is employed for occlusion changes and posture resistance [42]. They created a new network dubbed the regional attention network to adequately represent the significance of position variant FER and face areas in occlusion (RAN). Deep learning attention networks for the recognition of facial emotions [43], multi-attention networks for the recognition of facial expressions [44], and a new review on emotion recognition using facial appearance [45] are some of the related works for the recognition of FE. All the works mentioned [46] above have improved emotion recognition significantly over previous work. Still, none of these works contains a simple method for identifying essential face regions to detect emotions. This study [47] suggests a new framework based on a further attentiveness-coevolutionary neural network to focus on critical facial areas [48, 49].

The authors of [50] proposed ENCASE to combine expert features and DNN (Deep Neural Networks) for electrocardiogram (ECG) classification. They investigated specific

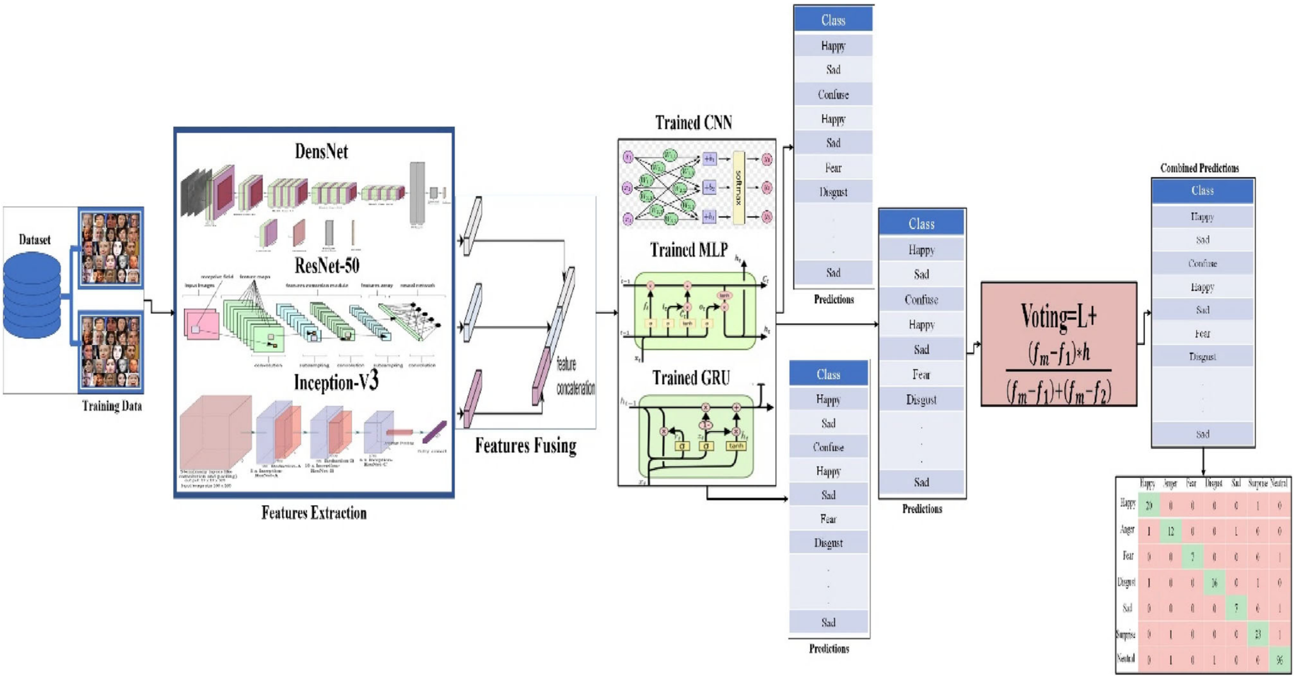
Table 1 Performance of deep learning multi-layer feature-fusion methods

Author name	Year	Methods/algorithms details	Dataset	Accuracy (%)
Bolioli [20]	2022	Inter-layer and intra-layer feature-fusion using InceptionV3 and VGG16 CNN architecture	UCM and NWPU	97.7% and 94.7%
Malakar [57]	2020	Multi-level convolutional neural network (MLCNN) approach by an ensemble of feature maps from different layers	FER-2013	73.03%
Bacanin [58]	2021	Hybrid multimodal (audio + Video feature data fusion)	AFEW	61.87%
Karras [25]	2022	Multi-feature fusion with an ensemble of CNN subnets approach	FER-2013	85.19%
Marzouk [38]	2022	Multi-region ensemble CNN (MRE-CNN) approach with feature fusion	AFEW and RAF-DB	47.43% and 76.7%
Dimitrios [42]	2019	Multi-feature fusion of temporal appearance features and temporal geometry features fusion of local and global	CK +	97.25%
Bhattacharya [46]	2022	Features with directional local binary pattern (DLBP) and discrete cosine transform (DCT) methods	CK	97%

applications and applied them to statistics, signal processing, and medicine. Then, they developed DNN for automatic deep feature extraction [51]. They also proposed a new algorithm to find the most representative wave (called the central wave) in long ECG recordings and extract the features of the central wave. Finally, they combined these features and subjected them to ensemble classification. The authors of [52, 53] proposed a new remote sensing scene classification approach using a set of robust and independently trained Deep Rule-Based (DRB) classifiers with different spatial information levels. Each DRB classifier is a comprehensive parallel set of transparent and human-interpretable zero-order fuzzy rules with a prototypical nature. The DRB classifier can be organized "from scratch" and built upon its structure. Using pre-trained neural networks as feature descriptors, the proposed DRB ensemble can show human-level performance through a parallel and transparent training process [54, 55]. Numerical examples on reference datasets show that the proposed method is more accurate when DRB classification creates fuzzy rules for human understanding [56–58]. According to the aforementioned literature review, numerous researchers are working to improve the performance of current CNN models on real-time facial expression and DFE datasets, which contain both real-world and lab-trained images. As a result, a powerful model with deep learning fusion techniques is required to accurately classify driver emotions and extract key image features. As a result, in our research, the author recommends the CNN-based DenseNet model. The most recent performance analysis based on recent research on multi-feature fusion-based methods using deep learning techniques is shown in Table 1.

The proposed methodology

The proposed ensemble learning framework involves several levels of merging different classifiers trained on different feature sets. Figure 2 illustrates the entire process of the proposed framework. Specifically, as shown in the feature preparation layer in Fig. 2, various feature sets are prepared using various feature extraction techniques. Furthermore, the feature set obtained by applying a specific method is further processed by applying various feature selection methods to obtain various feature subsets. Three state-of-the-art CNN models, DenseNet, InceptionV3, and Resnet-50 are used to extract high-dimensional features from the train and validation set images. The extracted fused features are concatenated to create a fused features vector. An ensemble classifier is developed using three state-of-the-art deep learning classifiers, CNN, gate recurrent unit (GRU), and multi-layer perception (MLP). The final output is selected by a voting scheme and assigned a class label predicted by most of the classifiers. In a real-world scenario, ensemble learning can be implemented in a more flexible configuration than that described in Fig. 2. For example, multiple base classifiers trained on the same feature set can be combined into a primary integration, then combined with the remaining base classifiers to form a quadratic integration. In this case, a secondary integration can be created from each feature set, and then some, all, or a secondary integration can be combined to create a top-level or even a final integration. The proposed framework can be viewed as a philosophical strategy for structural thinking and can also be used to solve the problem of driver emotion recognition. For example, multiple base classifiers trained on the same feature set can be combined into a primary integration, then combined with the remaining base classifiers to form a quadratic integration. In this case, a secondary integration can be created from each



	Happy	Anger	Fear	Disgust	Sad	Surprise	Neutral
Happy	20	0	0	0	0	1	0
Anger	1	12	0	0	1	0	0
Fear	0	0	7	0	0	0	1
Disgust	1	0	0	16	0	1	0
Sad	0	0	0	0	7	0	1
Surprise	0	1	0	0	0	23	1
Neutral	0	1	0	1	0	0	95

Fig. 2 Proposed ensemble classification model for driver emotions recognition

feature set, and then some, all, or a secondary integration can be combined to create a top-level or even a final integration. Granular computing can be viewed as a philosophical strategy for structural thinking, but it can also be used to solve the problem of driver emotion recognition. Each integration can be conceptualized as a single model in an ensemble learning framework because it contains multiple classifiers. Images can be of different sizes, which led to the development of the concept of particle size. Image size assists in improving the proposed model performance as it changes the proportion corresponding to the model size. The length and width of the images and models can vary considerably. Each level of classifier fusion can be interpreted as a different level of granularity in the proposed input learning framework.

Three models are used in this stage: GRU, MLP, and CNN. MLP is frequently used to solve problems requiring supervised learning and research into computational neuroscience and parallel distributed processing. The GRU is a type of RNN that uses less memory than long short-term memory and considers more efficient. However, using datasets with longer sequences improves the accuracy of LSTM. GRU occasionally has advantages over LSTM that are greater. These models ran on the data from the task of feature fusion and feature concatenation. Every model also has a distinct architecture where various tasks are carried out. The next step is to predict the data in tabular form so that a confusion matrix can be effectively generated after these models have been fully processed. The final stage will involve performing cure graphs in a simulation environment based on the confusion matrix.

In this stage, three models are used: Gate Recurrent Unit, Multi-Layer Perception (MLP), and CNN. MLP is commonly used to solve problems that require supervised learning and research in computational neuroscience and parallel distributed processing. The GRU is a type of RNN that uses less memory than LSTM and is considered more efficient. However, using datasets with longer sequences improves the accuracy of LSTM. GRU occasionally has advantages over LSTM that are greater. These models ran on the data from the task of feature fusion and feature concatenation. Each model also has a distinct architecture where various tasks are carried out. The next step is to predict the data in tabular form so that a confusion matrix can be effectively generated after these models have been fully processed. The final stage will involve performing cure graphs in a simulation environment based on the confusion matrix.

Several benchmark datasets, including AffectNet, CK+, FER-2013, JAFFE, and custom-developed datasets, are used to train the proposed ensemble CNN classifier. It should be noted that we trained a separate model for each dataset used in this study. A separate validation and test set are used for model parameter tuning and performance evaluation. A confusion matrix is an important metric consisting of various measures such as true positive, false positive, false negative, and true negative. These measures are used for validating model performance using the accuracy, precision, recall, and f1-score metrics.

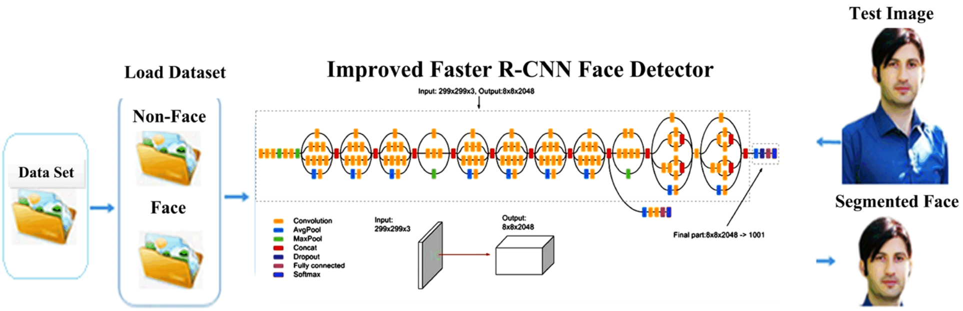


Fig. 3 DFER and emotion detection using an improved faster R-CNN model

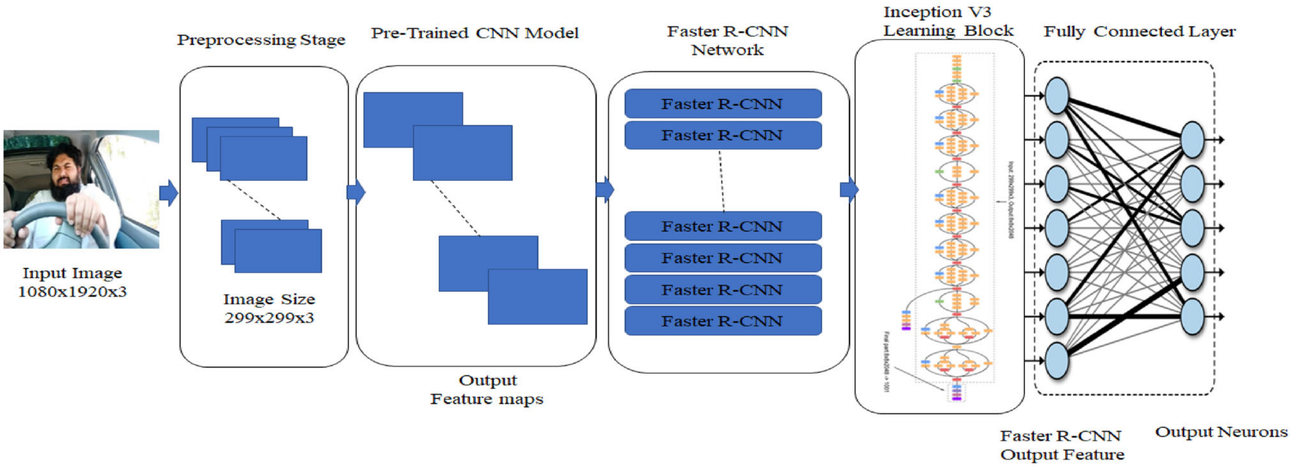


Fig. 4 Driver emotions recognition system CNN architecture

Face detection system

The face detection system allows images to pass through a specific process of facial recognition and be detected using the existing dataset. It combines R-CNN and deep learning methods for images with rectangular regions and schemes that utilize CNN features. The Faster R-CNN model can detect objects in two steps. The first step identifies a subset of regions on the target image where an object is present, while the second step is used to classify those objects in all regions, as shown in Fig. 3.

The proposed face detector is named "Improved Faster R-CNN," which uses the two-step scheme for emotion recognition. In the first step, fully connected layers are applied to the image in the detector with multiple layers up to the actual image and the segmented image. The image is placed in the system, where the dataset checks the image with the corresponding regions and other feature extractions. The position and expression of the image are matched with the dataset, and then the detector processes it further for analysis and recognition. In the second module, the improved Faster R-CNN is applied to the proposed areas of the images. Figure 3 shows a clear example of the face detection system where the two images show the actual image named as a test image and the other as a segmented image, the output image. A custom CNN block (Inception-V3) replaces the Improved Faster R-CNN feature-learning block to improve the accuracy and efficiency of face detection as shown in Fig. 4.

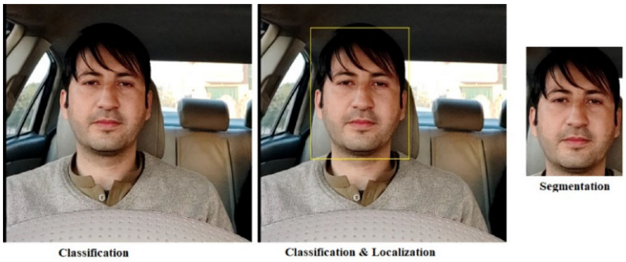


Fig. 5 Difference between classification, localization, and segmentation

Regional convolutional neural network (R-CNN)

In the image processing stage, the improved faster R-CNN algorithm uses edge boxes to generate regions of interest. The proposed scheme then resizes and crops the image to the appropriate size. The resized region is then processed by a CNN, which uses features trained by SVM to identify the size and shape of the image, as shown in Fig. 5.

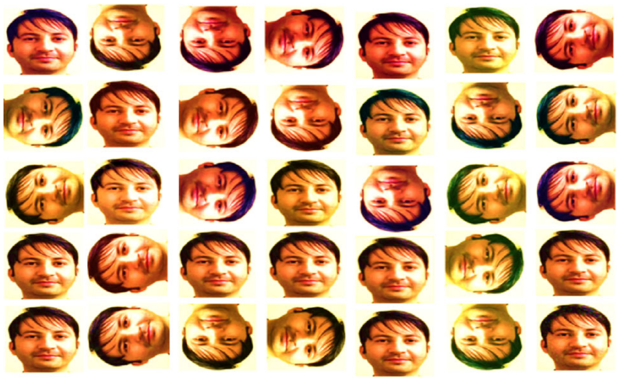


Fig. 6 Image augmentation using various image processing techniques

Data augmentation

In computer vision, data augmentation is used with different approaches, to increase the images in the given dataset by organizing and analyzing the existing dataset. Using image processing techniques, a single image is replicated to increase the quantity of image data that will be effective in computer vision and deep learning-based models for situations where the original dataset size is small. There are many ways to improve data, such as by changing the red, green blue (RGB) colors, using affine transformation, translating, rotating, adjusting the contrast, adding noise, taking noise away, changing the blueness, sharpening, flipping, cropping, and scaling, shown in Fig. 6.

Transfer learning-based on driver emotion recognition (DER)

Deep learning models are utilized to acquire transfer learning techniques to improve their performance when applying these methods to one-to-many challenges. This method has been used almost solely for object recognition in applications such as image speech recognition and image recognition in computer vision [49]. Other applications that have made use of this technology include. This method has been proposed to analyze and evaluate the vibrant images placed in the detector for later evaluation. This approach has been applied to the training of the benchmark dataset to use a strategy of test data and augmented data. When the entire dataset is about to be fully trained, the novel model must be used, whereas transfer learning does not require the model to be trained for many epochs. This approach can decrease the computation burden.

Transfer learning

Transfer learning is the most commonly used for the following processes and steps. Figure 7 depicts the preceding approach and steps of transfer learning, with each step elaborated differently.

The data are loaded into the pre-trained network at the initial step for further analysis. For the new task, the weight will be contingent upon the existence of data. After that, the final layers have to be loaded and replaced with the final layers in which fewer lanes are used to learn faster. Then comes the new layer with the train network phases in which the 100 s of images with 7 s of classes take place. Then, it proceeds to the next level, in which the prediction and

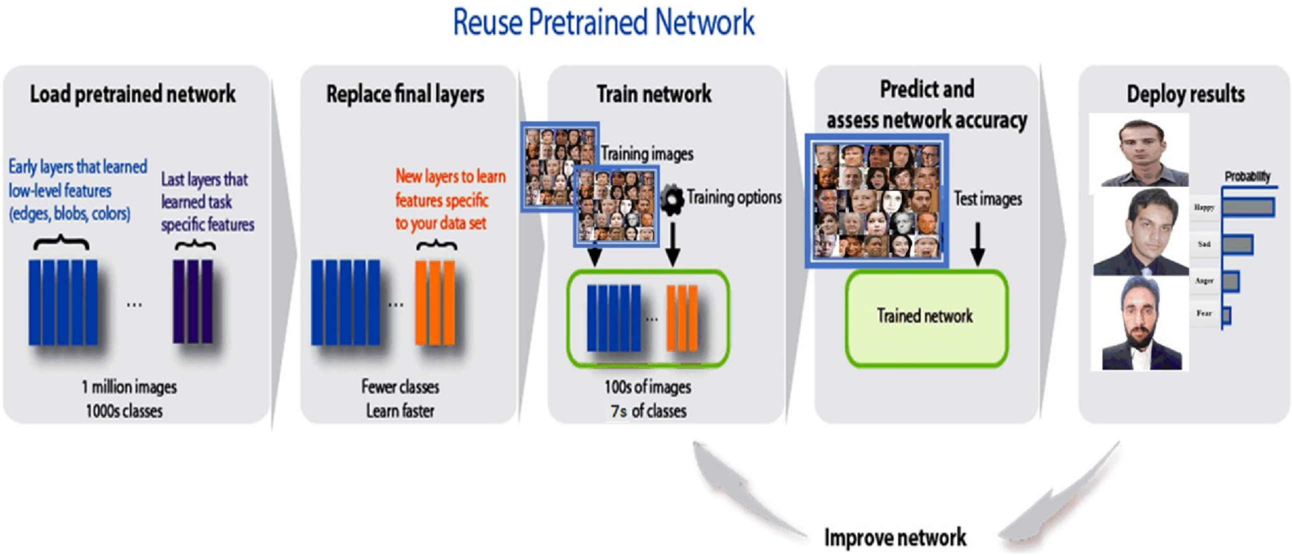


Fig. 7 Transfer learning workflow

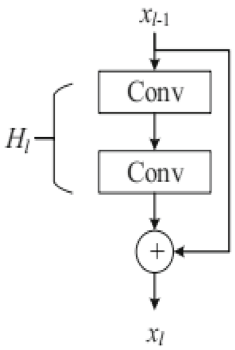


Fig. 8 Structure of the ResNet block

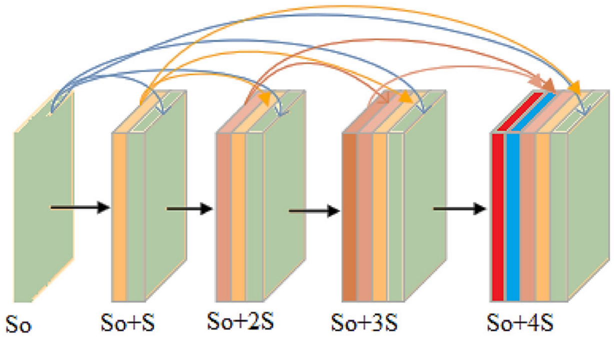


Fig. 9 Configuration of the DenseBlock

evaluation of the network for accuracy take place. In the last stage, the results have to be deployed in which the outputs are generated accordingly.

Transfer learning in pre-train CNNs

DenseNet

The ResNet structure is given in Fig. 8. In this scenario, the convolution layers are used by the CNN to be trained and [48] delivered to the next CNN stage with the increase of the values in Xi and Hi. In traditional CNN, all the layers are connected as given in the formula (1), which can go deep and make the network hard. In this terminology, it may come across as a gradient vanishing or exploding. After that, by skipping at least two layers, the ResNet offers an idea to be employed in some shortcut connections. With some transitions and conditions in which the input is whereas the output of the convolution layer is which is added with the shortcuts for the input layers, Therefore, the summative of the output is what is illustrated in Eq. 2.

After that, DenseNet can revise the model with the concatenation of the whole feature map accordingly. The expressions are given in Eq. 3, in which the feature maps from the past layers are instead of a summation of the output:

$$x_i = H_i(x_{i-1}) \tag{1}$$

$$x_i = H_i(x_{i-1}) + x_{i-1} \tag{2}$$

$$x_i = H_i([x_0, x_1, x_2, \ldots, x_{i-1}]) \tag{3}$$

where i stands for the index of layer, H denotes the operation of non-linear, and X_i expresses the features of the lth layer.

The block diagram of the DenseNet is given in Fig. 9. With the focus on Eq. 3, the DenseNet can offer the concatenation of the previous maps to the previous layers. This terminology states that the maps of features are gathered and directed to one newly generated feature map. The DenseNet, which is newly designed, can propose advantages such as gradient vanishing to decrease the problem with the exploding manner, reuse, Etc. However, for the structure of the DenseNet to become feasible, some of the following changes need to be made. In which the downsampling is used to create a possible concatenation. The total given steps are given in Fig. 9, in which, from left to right, it precedes and increases with the S + 1. S is the actual module and the other values added with the module, such as S + 2, S + 2, and S + 4, are the concatenation maps.

Each layer is linked with another as it makes a total combination of 10 transitions. Figure 10 shows that each map for generating the S features is included with one of the operations of H1. The total of 5 layers in which the Sth values are introduced with each layer from the highest of S0 + 4S, in which the term So denotes the number of features mapped with the layer of the previous one. In this study, 32 has been kept as the default value of S. Nevertheless, there is a massive number of inputs to the networks, in which a layer named "bottleneck" was also introduced for the DenseNet. The convolution layer designated that layer before the layer of convolution. That layer has helped decrease the number of image features with the solution of the cost of computation. After that, considering the model's accuracy, a layer named "transition" is used to reduce the feature maps in the given procedure. The supposition is given if the S feature maps are to be generated with the Dense Block with the assumption of the value of the compression factor. Therefore, the maps of features will have to be minimized. If the value is so, the number of features on the maps will have to be the same. Figure 10 shows the relation between Dense Blocks and Transition Layer.

The overall structure and procedure of the DenseNet have been illustrated in Fig. 10, in which the input layers, GAP layers, Dense Blocks, and transition layers have been given with transitions from one layer to another. With the normalized batch layer, the transition layers consist of these with the value of convolution layers and the middle layers of pooling

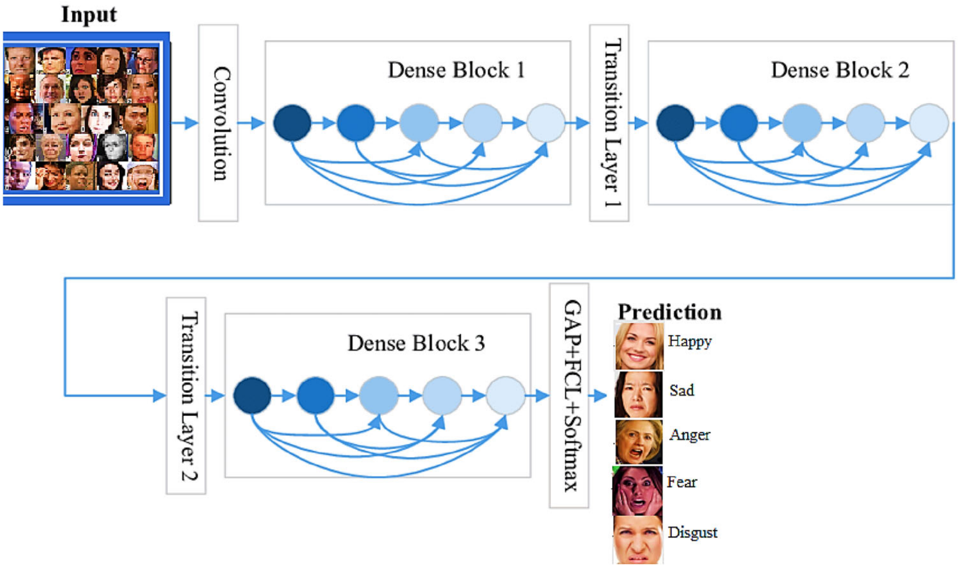


Fig. 10 Dense Blocks with relation to the Transition Layer

in which the two are kept in stride. In particular, the value of GAP is identical to the traditional pooling methods, but the term GAP has to undergo more powerful features in the aspect of reduction, which can reduce the map features by the value. It denotes that the term GAP layers are being reduced to a single digit as a whole slice.

Features weights optimization

For optimization of the detailed model to train our dataset for fine-tuning and to pre-train CNNs, we have separately evaluated each one. From the result, we have used the DenseNet pre-train CNNs. The study we propose has the uniqueness to augment the image. Furthermore, the additional dataset for the training has been generated with the usage of the technique of augmentation. To mitigate the overfitting problem, the data augmentation can be trained during the training with the collaboration of CNNs models. We have applied some randomized vertical and horizontal shifts with the extent of the 10% to the originality dimension in the study. By doing so, further, the rotation of randomized was fit to 20% which was applied for the images to train with having a small zoom of random.

In addition, we rotate the images horizontally to enhance dataset size. We eliminated all fully connected layers to optimize each network and utilized only the convolutional part of each model's architecture. At the final stage of the convolutional layer, we add a global mean polarity layer, followed by a classification layer with SoftMax nonlinearity. Using a learning rate of 0.0001 and a speed of 0.90, 50 iterations of stochastic gradient descent (SGD) optimization are used to refine the network. In all circumstances, the loss function is equal to the cross-entropy squared. It is used to alter the validation set's hyperparameters. To clarify, each network's input has a distinct shape. The initial stage in data preparation entails resizing all photos based on the model input and saving them in many files of various formats. The same initialization and learning rate rules are used to train both models.

Table 2 Specification of GPU used for model training

Manufacturer	Nvidia
Model model	$RTX2080Ti$
Memory	$4GBGDDR - 4$
Cores	4352
TMUS	272
ROPS	88
Buswidth	352bits

Results and discussion

To proposed driver facial expression recognition (DFER) method has been effectively tested and proven on various standard datasets for the development of this section. In this study, we compared the current FER method with the state-of-the-art method, and the data collected as also encompassed the result of the qualitative and quantitative evaluation. The proposed system used two reference datasets. Every dataset is divided into two sections such as training, and validation sets randomly. The training sets will be 70% of the datasets and 30% will be for validation sets of each dataset. All simulations are performed under the simulation environment using the MATLAB R2021a platform contained in the proposal. All of this is done on a workstation shown in Table 2.

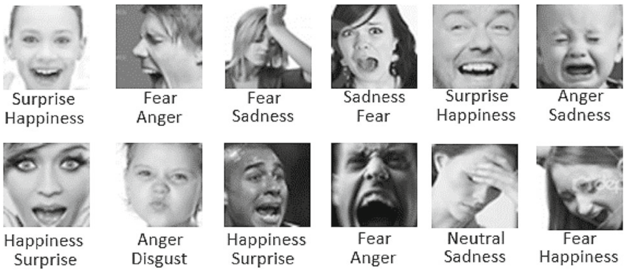


Fig. 11 Some random images from the FER-2013 dataset

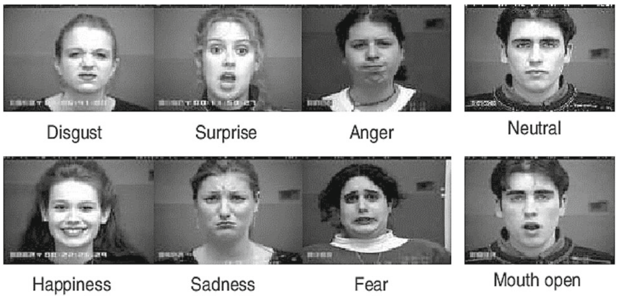


Fig. 12 Seven images from the CK+ dataset [45]

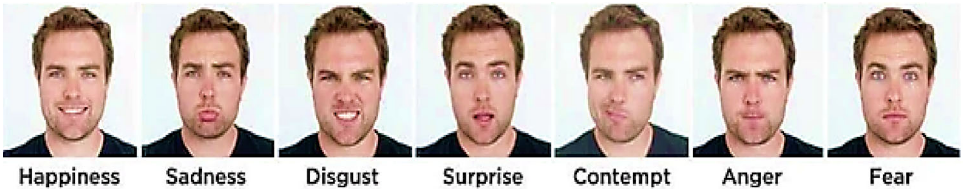
Datasets

Due to funding constraints, workload constraints, time constraints, and algorithm performance evaluation requirements, most FER researchers rely on benchmark datasets. The most commonly used normative datasets range from sentiment inquiry to evaluation. The benchmark datasets are the extended Cohn–Kanade (CK+), Japanese Female Facial Expressions (JAFFE), and FER-2013. In this work, we used the FER-2013 facial expression dataset [5], CK+ [49], AffectNet [49], JAFFE [14], and a custom-developed dataset for DFER, which are among the benchmark datasets used for DFER. This section will briefly overview the benchmark and custom datasets used in this work. After that, it will provide the performance of our models on benchmark datasets along with a custom-developed dataset and compare the results with some of the current sound work.

FER-2013 dataset

The ICML-2013 was the first dataset used to represent the data for emotion recognition based on the existing dataset [5]. The FER-2013 dataset consists of different images used to analyze and evaluate the proposed work. A total of 35,887 images were included in this dataset. In which the 48/48 resolution was set. The majority of the images were taken in the real-life scenario field. There are a total of 28,709 images these images are for the training set, and the 3589 images are for the test set. With the Google Application Programming Interface (API), the automatically captured datasets can be retrieved from the Google Image Search. The essential aspects of facial expression recognition are the sixth or neutral expressions to be applied to the faces. The dataset named FER-2013 is a common aspect of face recognition that shows low contrast, and facial occlusion is made in additional datasets. From this dataset, some pictures are given in Fig. 11.

Fig. 13 JAFFE dataset [46]



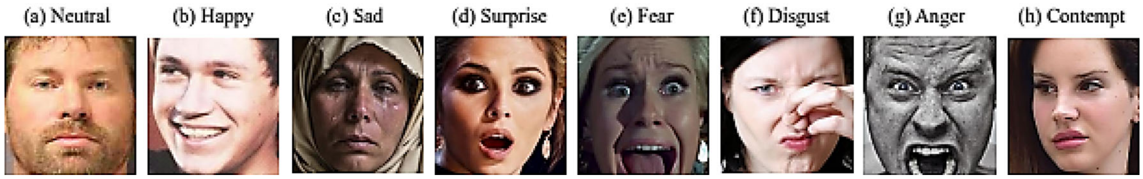
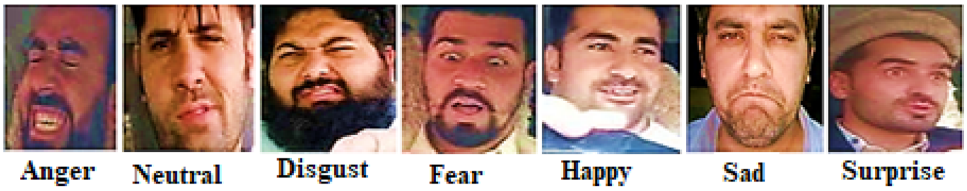


Fig. 14 Images from the AffectNet dataset [47]

Fig. 15 Images from the custom-developed dataset [48]



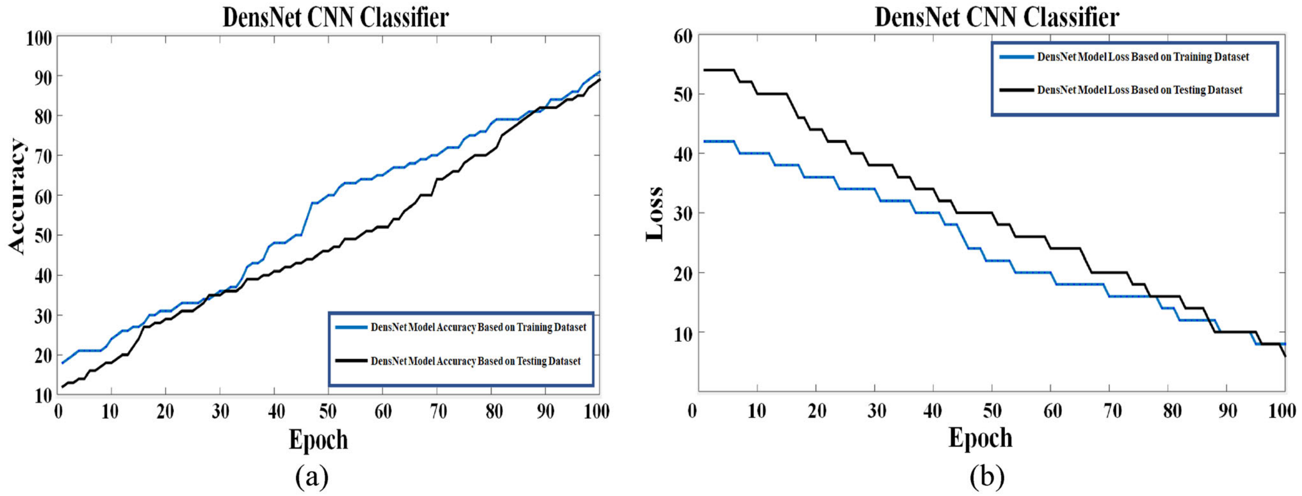


Fig. 16 DeneseNet CNN models based on training and validation, **a** accuracy and **b** loss plot for JAFFE original datasets

Table 3 Detailed test accuracy by class of JAFFE original dataset

Class	Accuracy	Precision	Recall	F-measure
Happy	100	100	100	100
Anger	100	100	100	100
Fear	95	75	100	84
Disgust	100	100	100	100
Sad	68	59	39	64
Surprise	79	100	44	45
Neutral	100	100	100	100
Average	91.71	90.57	83.28	91.28

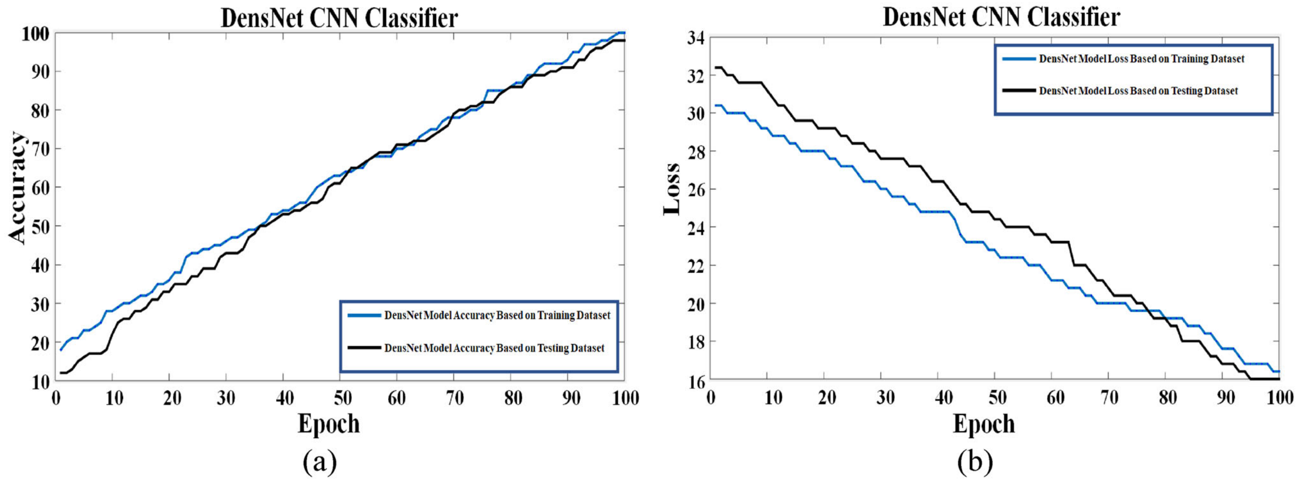


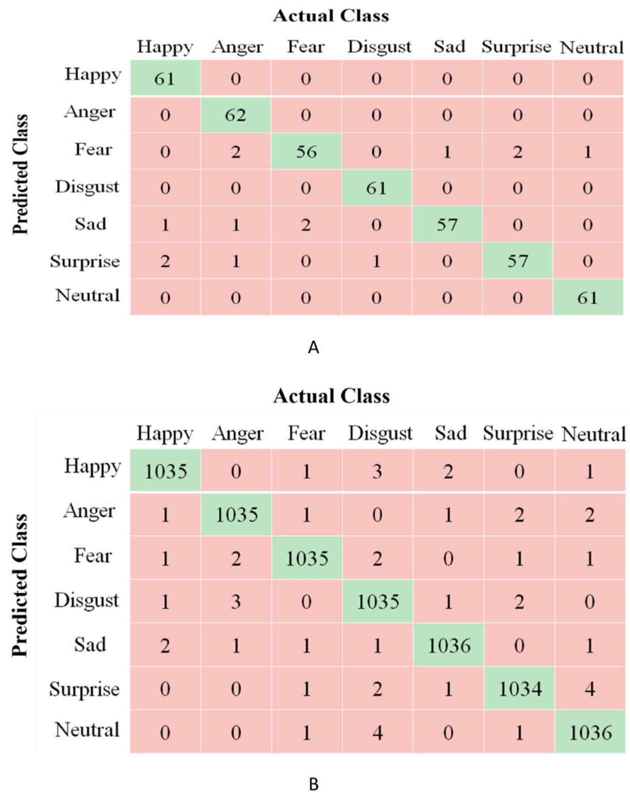
Fig. 17 DeneseNet CNN model based on training and validation, **a** accuracy and **b** loss plot for JAFFE-augmented dataset

CK + dataset

The dataset named CK+, which is Cohn–Kanade, is a dataset for facial expression images [45]. It is a publicly available dataset for recognizing the driver's facial expressions as an active unit. It has non-posed and posed expressions in which the analysis can be made with ease. In this dataset, the overall number of images was 593 in an aspect of sequence across the number of 123 subjects. This sequence's last frame was taken from the existing works used for the FER image base. For a

Table 4 Detailed test accuracy by class of augmented JAFFE dataset

Class	Accuracy	Precision	Recall	F-measure
Happy	98.99	96.98	98.15	96.97
Anger	97.96	97.99	96.96	97.19
Fear	98.48	96.45	99.92	97.18
Disgust	99.45	98.95	98.93	98.44
Sad	97.78	97.12	97.01	97.10
Surprise	98.49	99.97	98.11	94.17
Neutral	99.49	98.98	98.02	98.90
Average	98.66	98.06	98.15	97.13



Actual Class

Predicted Class	Happy	Anger	Fear	Disgust	Sad	Surprise	Neutral
Happy	61	0	0	0	0	0	0
Anger	0	62	0	0	0	0	0
Fear	0	2	56	0	1	2	1
Disgust	0	0	0	61	0	0	0
Sad	1	1	2	0	57	0	0
Surprise	2	1	0	1	0	57	0
Neutral	0	0	0	0	0	0	61

A

Actual Class

Predicted Class	Happy	Anger	Fear	Disgust	Sad	Surprise	Neutral
Happy	1035	0	1	3	2	0	1
Anger	1	1035	1	0	1	2	2
Fear	1	2	1035	2	0	1	1
Disgust	1	3	0	1035	1	2	0
Sad	2	1	1	1	1036	0	1
Surprise	0	0	1	2	1	1034	4
Neutral	0	0	1	4	0	1	1036

B

Fig. 18 Confusion matrix of JAFFE **a** original and **b** augmented dataset

total of seven images, samples are given from this dataset in Fig. 12.

JAFFE dataset

The JAFFE dataset [59] expresses basic FE in Japanese models (female). This dataset contains two hundred and thirteen (213) images of seven different FE. With the addition of the contempt class in CK+, each dataset has to use seven basic FE, which are most commonly used. To rate each image, 60 Japanese subjects were rated using 7 FE [46]. Figure 13 shows seven images from the dataset.

AffectNet dataset

AffectNet is one of the largest freely available datasets in the FER work [47]. AffectNet is a new real-life FE dataset consisting of FE and annotations. AffectNet is an FE dataset of over 1 million facial images collected from the internet. There are over 1250 emotions from 6 different countries, people with 3 main search engines—the presence of seven different EFs (deterministic model) restored images (440,000, dimensional model). AffectNet is the most publicly available EF, valence, and pacing dataset, enabling investigators to conduct investigations automatically. FER in two distinct emotional models. There are two main lines in the hierarchical model. A deep neural network is used to classify the images and predict the symmetry and intensity of the simulation. Accuracy is based on seven categories (happy, surprised, sad, fear, disgust, anger, contempt, and neutral) in Fig. 14.

Custom developed dataset (CDD)

In custom datasets, the own mind datasets are created and used for analysis and evaluation in which the proposed and the existing datasets are combined for the FE recognition of the driver [48] in Fig. 15. The deep learning approach filters and extracts the features from the given custom and state-of-the-art datasets. In the moving of a vehicle with a duration of exceeding time, these images have used the expressions of the driver with the real-time scenario to enhance and capture the right moment. Each image with the subject will be tested against the extracted features. Multiple images are taken from moving vehicles such as the Toyota Land cruiser, Honda Civic, and Toyota Prius. From these scenarios, every subject's data are recorded for up to 10 min each. Every subject in this study is a driver male with age 25 to 40 in which some of them have a beard or no beard, and wear a cap or no cap. As well as the videos were also recorded and evaluated for recognition in real time in which the obstacles come in a way, and the light changes as the vehicle moves forwards for the emotion recognition system. For the analysis and evaluation of the DFER, the proposed dataset and benchmarks

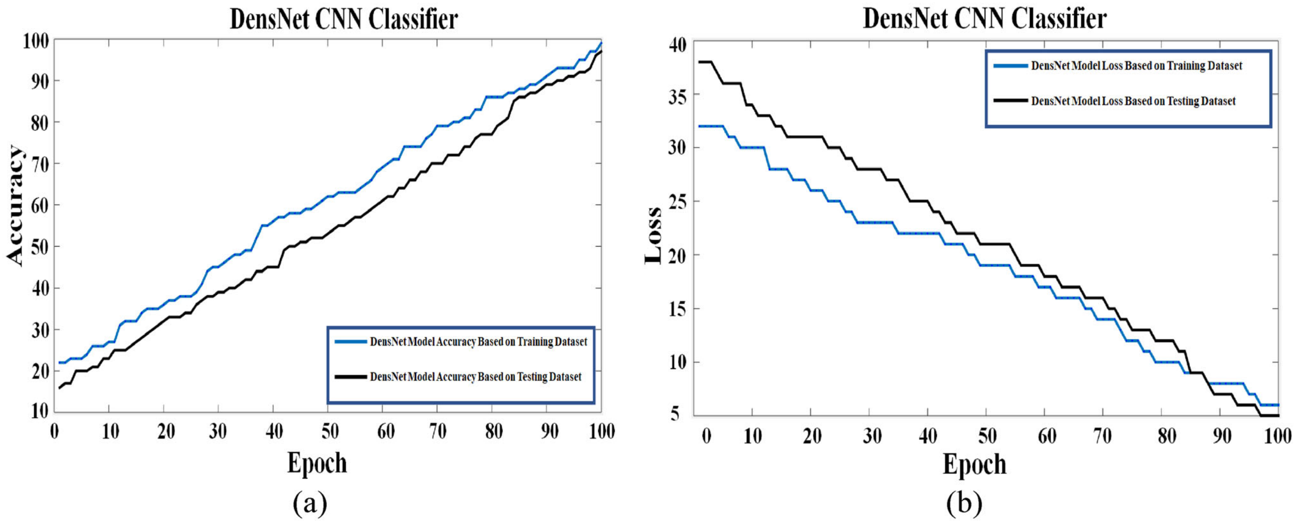


Fig. 19 DenseNet CNN model based on training and validation, **a** accuracy and **b** loss plot for CK+ original dataset

Table 5 Detailed test accuracy by class of CK+ original dataset

Class	Accuracy	Precision	Recall	F-measure
Happy	98.45	98.55	99	96.29
Anger	99.85	96.15	98	96.77
Fear	98.24	100	98.99	89.33
Disgust	98.07	96.18	99	95.42
Sad	99.99	89	99	96
Surprise	98.7	99.25	99.01	98.15
Neutral	99.89	97.95	96.96	89
Average	99.02	96.72	98.56	94.42

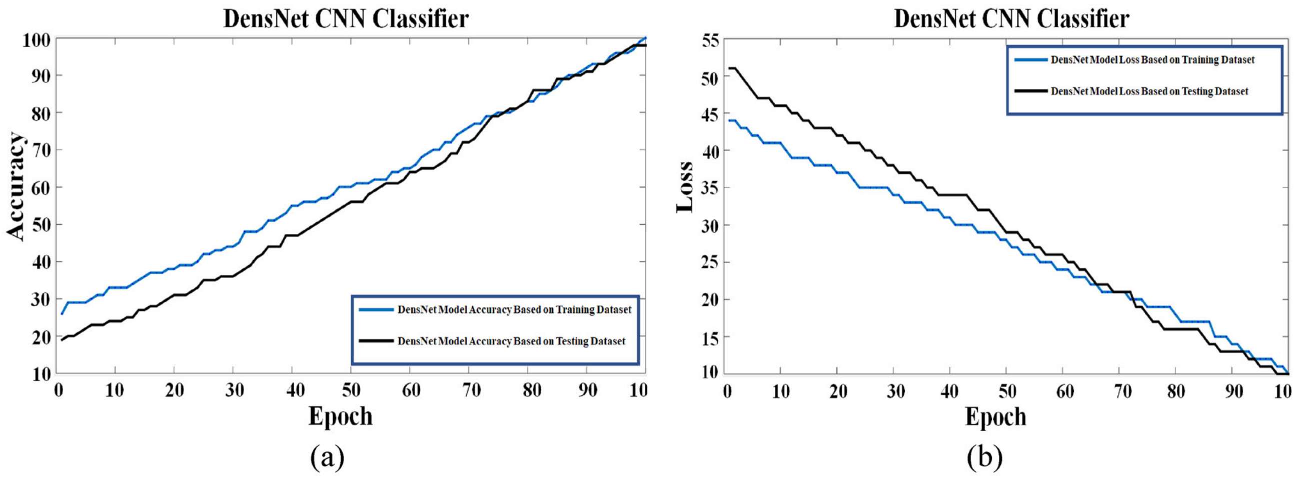


Fig. 20 DenseNet CNN model based on training and validation, **a** accuracy and **b** loss plot for CK + augmented dataset

have been trained under the deep learning models with the custom-made dataset.

Results

The DFER method described above was tested and found useful on several standard datasets used in the development of

Table 6 Detailed test accuracy by class of augmented CK+ dataset

Class	Accuracy	Precision	Recall	F-measure
Happy	99.84	99.56	98.99	99.28
Anger	99.24	99.13	98.92	99.02
Fear	99.85	98.48	98.48	98.48
Disgust	99.92	99.66	99.66	99.66
Sad	99.6	96.64	94.18	95.39
Surprise	99.81	99.51	93	99.27
Neutral	99.84	94	99.84	99.84
Average	99.72	98.14	97.58	98.7

this section. This study also includes a quantitative and qualitative evaluation of valid measures of outcomes obtained from data collection and a comparison of the proposed technique with current FER techniques. A more concrete example is that the proposed system uses five reference datasets. Each dataset in the proposed system is randomly split into training sets and test sets, and the training set is much larger than the test set.

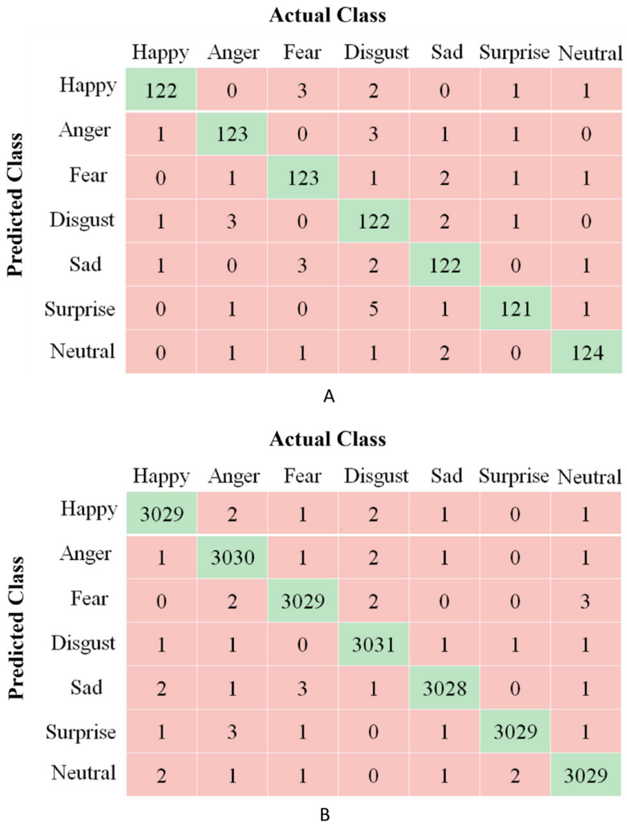
Experiments on the JAFFE dataset

Experiments on the JAFFE dataset are conducted using a random hold-out splitting strategy, which yields the most accurate results. In the first part, 189 images (70%) are used for training and 24 images (30%) are used for validation. In the second stage, 6237 training images are added to the JAFFE-augmented dataset and used to support the model. 792 images were also used during model validation. Figure 16a shows an accuracy of 89.2% using the original JAFFE dataset, while Fig. 16b shows the DenseNet CNN model validation and training loss plots for the JAFFE dataset original. The confusion matrices of the original JAFFE dataset and the augmented dataset are shown in Fig. 18a and b, while the detailed classification test accuracy of the original JAFFE dataset is shown in Table 3.

Using the JAFFE-augmented dataset, the accuracy using the DenseNet CNN model is 98.01%, as shown in Fig. 17a, while Fig. 17b illustrates the training and model validation loss graphs DenseNet CNN for the JAFFE-augmented dataset. Figure 17a and b illustrates the comparative accuracy and loss of the proposed DenseNet CNN model for the augmented dataset is proportional to the number of epochs, while the detailed accuracy of the category test for the JAFFE-augmented dataset is shown in Table 4 (Fig. 18).

Experiments on the CK + dataset

Figure 19a and b shows the accuracy and loss of the training and validation of the original CK+ datasets. These studies were carried out using the random hold-out splitting method.



Actual Class

Predicted Class

	Happy	Anger	Fear	Disgust	Sad	Surprise	Neutral
Happy	122	0	3	2	0	1	1
Anger	1	123	0	3	1	1	0
Fear	0	1	123	1	2	1	1
Disgust	1	3	0	122	2	1	0
Sad	1	0	3	2	122	0	1
Surprise	0	1	0	5	1	121	1
Neutral	0	1	1	1	2	0	124

A

Actual Class

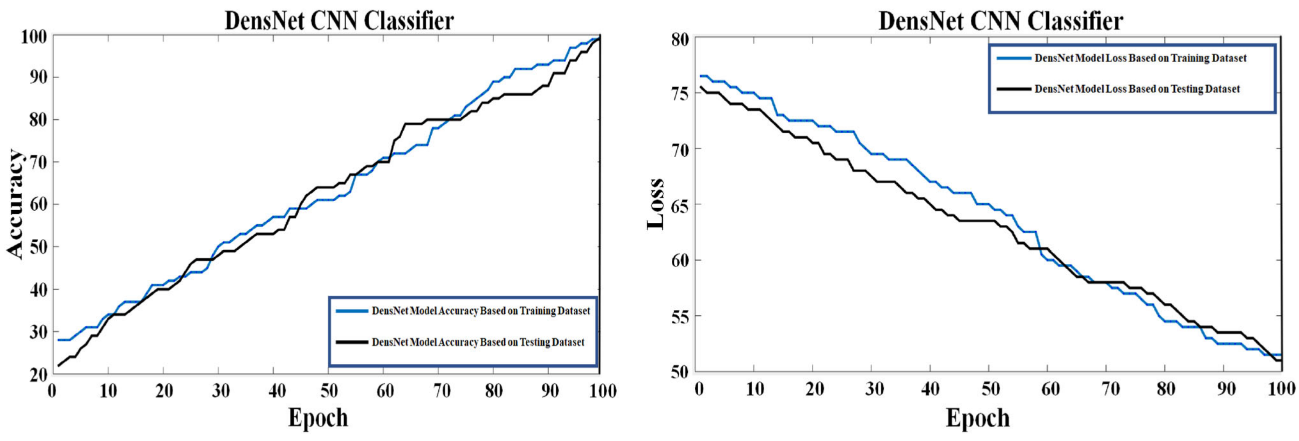
Predicted Class

	Happy	Anger	Fear	Disgust	Sad	Surprise	Neutral
Happy	3029	2	1	2	1	0	1
Anger	1	3030	1	2	1	0	1
Fear	0	2	3029	2	0	0	3
Disgust	1	1	0	3031	1	1	1
Sad	2	1	3	1	3028	0	1
Surprise	1	3	1	0	1	3029	1
Neutral	2	1	1	0	1	2	3029

B

Fig. 21 Confusion matrix of CK+ **a** original and **b** augmented dataset

There are two phases, with the first phase using for original CK+ dataset, 444 images 70% of the total for the training sets, using 192 images 30% of the total, for the validation sets. In the second phase, for the augmented CK+ dataset, a total of 14,652 training images and 6,336 validation images are taken. Figure 19a and b illustrates the model training and validation loss plot based on the CK+ original dataset to obtain an accuracy of 97.20%. The confusion matrix of the CK + original dataset and augmented dataset is shown in Fig. 21a and b while detailed test accuracy by class of CK+ Original dataset is shown in Table 5.

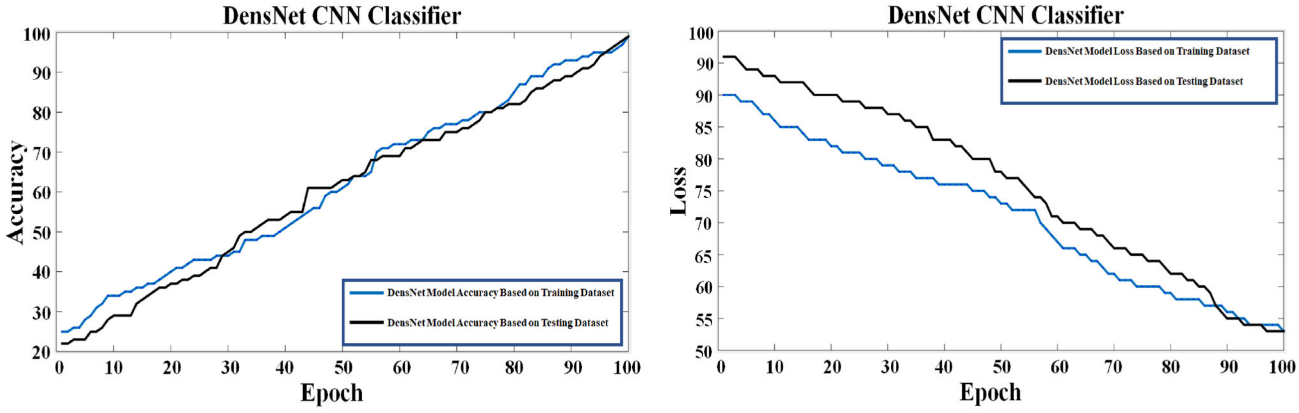


(a) (b)

Fig. 22 DenseNet CNN model based on training and validation, **a** accuracy and **b** loss plot for FER-2013 dataset

Table 7 Detailed test accuracy by class of FER-2013 original dataset

Class	Accuracy	Precision	Recall	F-measure
Happy	99	98.21	99.85	98.03
Anger	99.1	99	99.56	94.36
Fear	99.97	99.25	99.5	99.45
Disgust	99.78	99.97	99.25	96.25
Sad	99.98	98.99	99.48	99.15
Surprise	99.58	99.89	99.75	99.78
Neutral	99.9	99.58	99.49	96.99
Average	99.61	99.27	99.55	97.71



(a) (b)

Fig. 23 DenseNet CNN model based on training and validation, **a** accuracy and **b** loss plot for FER-2013 augmented dataset

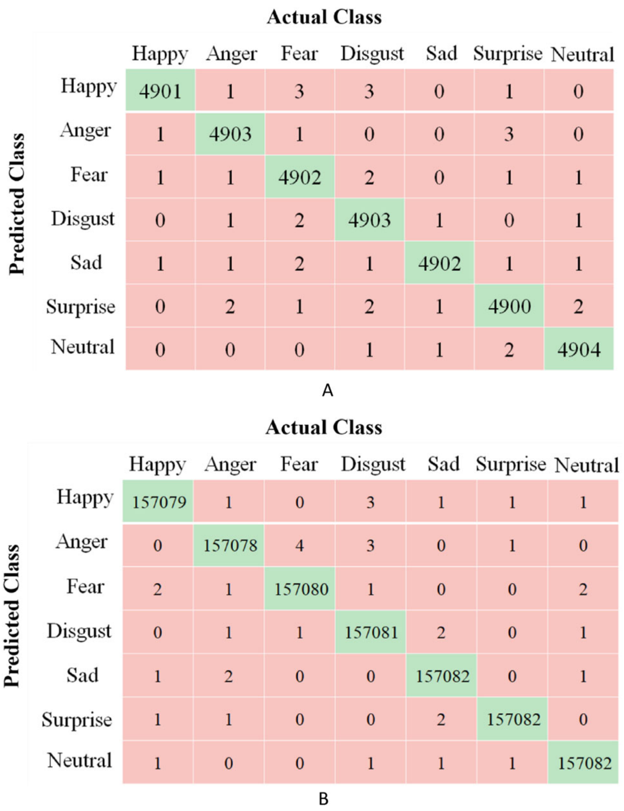
Figure 20a demonstrates that using the augmented CK+ dataset yields an accuracy of 98.53%, and Fig. 20b illustrates the model training and validation loss plot for the augmented CK+ dataset using the DenseNet CNN model while detailed test accuracy by class of CK+ augmented dataset is shown in Table 6 (Fig. 21).

Experiments on the FER-2013 dataset

Figure 22a and b shows the accuracy and loss of the original FER-2013 datasets for training and validation. These studies were conducted by randomizing hold-out split. There are two phases. The first phase was used for the original FER-2013 dataset, and the second phase was used for the

Table 8 Detailed test accuracy by class of augmented AffectNet dataset

Class	Accuracy	Precision	Recall	F-measure
Happy	89.59	98.39	99.65	90.48
Anger	98.96	99.38	90.65	99.36
Fear	99.87	99.01	96.76	91.89
Disgust	98.72	94.99	89.05	91.2
Sad	99.99	98.34	98.25	98.23
Surprise	99.88	96.37	97.7	99.36
Neutral	99.35	99.38	99.32	96.6
Average	98.05	97.98	95.91	95.3



Actual Class

Predicted Class	Happy	Anger	Fear	Disgust	Sad	Surprise	Neutral
Happy	4901	1	3	3	0	1	0
Anger	1	4903	1	0	0	3	0
Fear	1	1	4902	2	0	1	1
Disgust	0	1	2	4903	1	0	1
Sad	1	1	2	1	4902	1	1
Surprise	0	2	1	2	1	4900	2
Neutral	0	0	0	1	1	2	4904

A

Actual Class

Predicted Class	Happy	Anger	Fear	Disgust	Sad	Surprise	Neutral
Happy	157079	1	0	3	1	1	1
Anger	0	157078	4	3	0	1	0
Fear	2	1	157080	1	0	0	2
Disgust	0	1	1	157081	2	0	1
Sad	1	2	0	0	157082	0	1
Surprise	1	1	0	0	2	157082	0
Neutral	1	0	0	1	1	1	157082

B

Fig. 24 Confusion matrix of FER-2013 **a** original and **b** augmented dataset

augmented FER-2013 dataset. 23,569 images 70% of the total, for the training sets, and 10,658 images 30% of the total for the validation sets. In the second phase, for the augmented FER-2013 dataset used 777,777 training images and 321,691 validation images were collected. Figure 22a and b illustrates the model training and validation loss plot of the original FER-2013 dataset by the DenseNet CNN model. The overall accuracy of 99.01% was achieved by utilizing the original FER-2013 dataset. The confusion matrix of the FER-2013 original dataset and augmented dataset is shown in Fig. 24a and b, while detailed test accuracy by class of the FER-2013 original dataset is shown in Table 7.

99% accuracy is achieved using the FER-2013 augmented dataset, as shown in Fig. 23a, while Fig. 23b shows the validation loss plot of the augmented FER-2013 dataset using the DenseNet CNN model while detailed test accuracy by class of FER-2013 augmented dataset is shown in Table 8 (Fig. 24).

Experiments on the AffectNet dataset

Figure 25a and b illustrates the training and validation accuracy and loss of the original AffectNet dataset. These studies were conducted by randomizing the hold-out split. The first phase uses 187,807 images which are 70% of the total for the training sets, while to use of 87,346 images which are 30% of the total for the original AffectNet dataset, as validation images. In the second phase, 2,817,105 training images and 1,310,190 validation images were used for the augmented AffectNet dataset. Figure 25a demonstrates an accuracy of 87.37%, while Fig. 25b shows the training and validation loss plot for the original AffectNet dataset by the DenseNet CNN model. The confusion matrix of the AffectNet original dataset and augmented dataset is shown in Fig. 27a and b while detailed test accuracy by class of AffectNet original dataset is shown in Table 9.

The accuracy achieved by the augmented AffectNet dataset is 96.81% as shown in Fig. 26a and b show the training and validation loss plot by the DenseNet CNN model while detailed test accuracy by class of AffectNet augmented dataset is shown in Table 10 (Fig. 27).

Experiments on the custom-developed dataset (CDD)

These studies were carried out using the randomized hold-out splitting method. Furthermore, Fig. 28a and b illustrates the accuracy and loss of the training and validation of the original CDD, respectively. In the first phase, 763,880 images 70% of the total for the training sets are used while the remaining 329,926 images representing 30% of the original CDD are used as validation images. In the second phase, 5,347,160

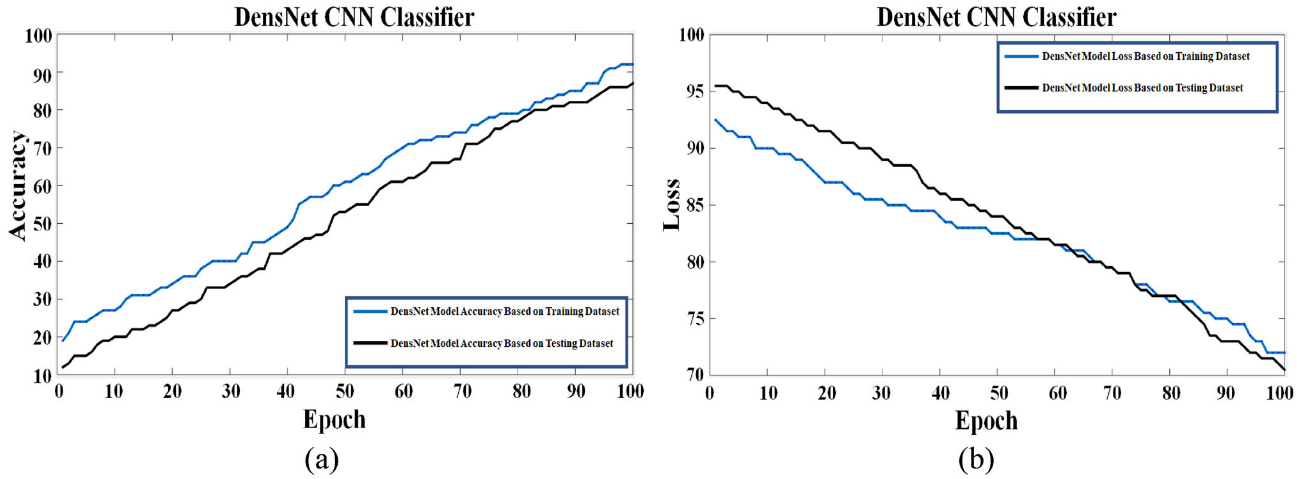


Fig. 25 DenseNet CNN model based on training and validation, **a** accuracy and **b** loss plot for AffectNet dataset

Table 9 Detailed test accuracy by class of AffectNet original dataset

Class	Accuracy	Precision	Recall	F-measure
Happy	86.78	82	80.56	90.48
Anger	90.65	72.56	89.42	80.75
Fear	96.76	89.87	74.94	64.26
Disgust	89.05	90.25	90.65	86.89
Sad	98.25	97.02	90.99	85.66
Surprise	97.7	92.71	80.38	72.89
Neutral	96.41	89.42	92.87	96.6
Average	93.65	87.69	85.68	82.5

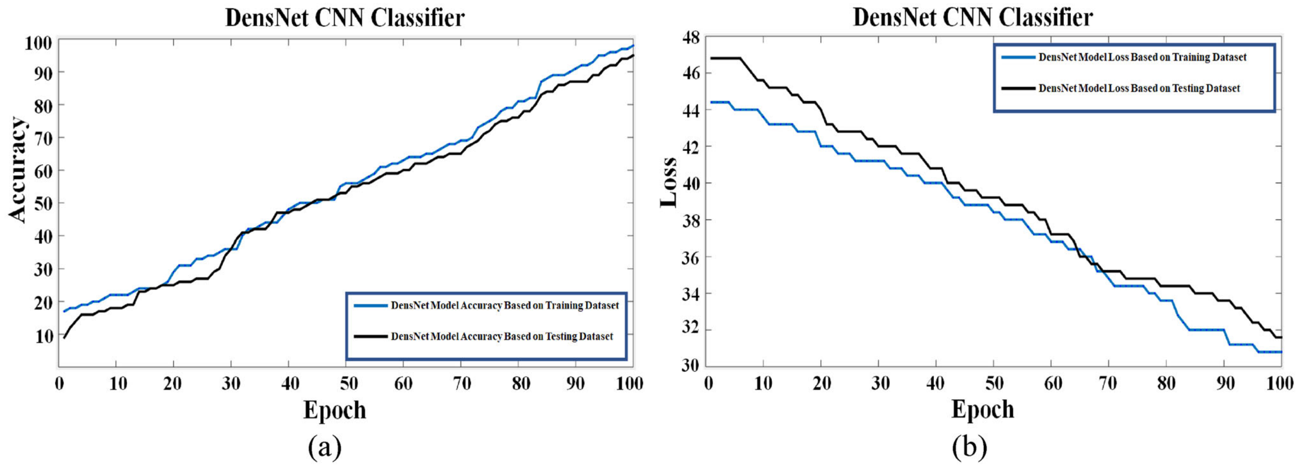


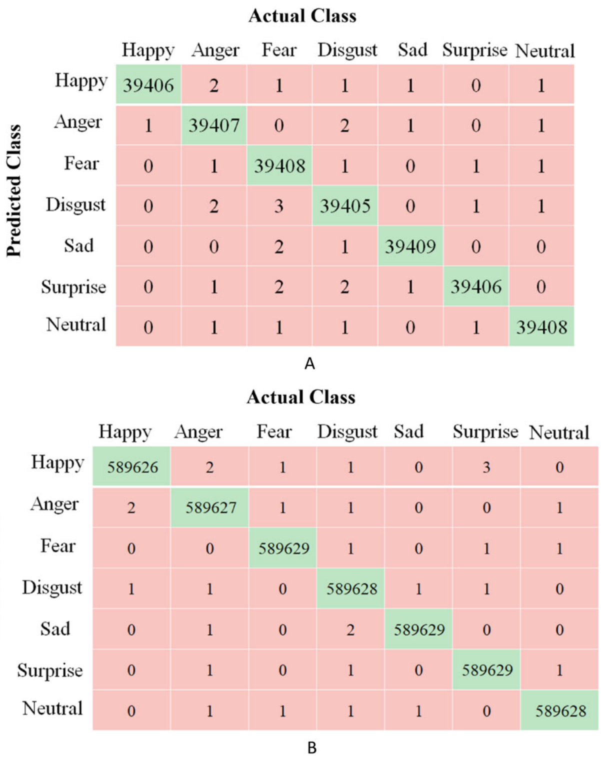
Fig. 26 DenseNet CNN model based on training and validation, **a** accuracy and **b** loss plot for AffectNet augmented dataset

training images and 2,309,482 validation images were collected for the augmented CDD. Figure 28a and b shows the model training and validation accuracy and loss plot for the CDD original dataset, which obtained an accuracy of 98.61% using the DenseNet CNN model. The confusion matrix of the CDD original and augmented dataset is shown in Fig. 30a and b while detailed test accuracy by class of CDD original is shown in Table 11.

The DenseNet CNN model achieves 99.90% accuracy on the augmented CDD, as shown in Fig. 29a and b the training and validation accuracy and loss plots while detailed test

Table 10 Detailed test accuracy by class of AffectNet original dataset

Class	Accuracy	Precision	Recall	F-measure
Happy	86.78	82	80.56	90.48
Anger	90.65	72.56	89.42	80.75
Fear	96.76	89.87	74.94	64.26
Disgust	89.05	90.25	90.65	86.89
Sad	98.25	97.02	90.99	85.66
Surprise	97.7	92.71	80.38	72.89
Neutral	96.41	89.42	92.87	96.6
Average	93.65	87.69	85.68	82.5


Actual Class
Predicted Class
A
Actual Class
Predicted Class
B

	Happy	Anger	Fear	Disgust	Sad	Surprise	Neutral
Happy	39406	2	1	1	1	0	1
Anger	1	39407	0	2	1	0	1
Fear	0	1	39408	1	0	1	1
Disgust	0	2	3	39405	0	1	1
Sad	0	0	2	1	39409	0	0
Surprise	0	1	2	2	1	39406	0
Neutral	0	1	1	1	0	1	39408

	Happy	Anger	Fear	Disgust	Sad	Surprise	Neutral
Happy	589626	2	1	1	0	3	0
Anger	2	589627	1	1	0	0	1
Fear	0	0	589629	1	0	1	1
Disgust	1	1	0	589628	1	1	0
Sad	0	1	0	2	589629	0	0
Surprise	0	1	0	1	0	589629	1
Neutral	0	1	1	1	1	0	589628

Fig. 27 Confusion matrix of AffectNet **a** original and **b** augmented dataset

accuracy by class of CDD augmented is shown in Table 12 (Fig. 30).

Statistical tests of the DenseNet model based on the aforementioned datasets

Descriptive statistics are used to summarize a set of observations, in order to communicate the largest amount of information as simply as possible. Statisticians commonly try to describe the following observations.

1. a measure of location, such as the arithmetic mean.
2. a measure of statistical dispersion.
3. a measure of the shape of the distribution like skewness or kurtosis

The ANOVA test of the DenseNet model is shown in Fig. 31 while multiple comparison results are presented in Tables 13 and 14. The DenseNet model using original datasets is shown in Fig. 32 while augmented datasets are shown in Fig. 33.

Each cell of the matrix is weighted by its proximity to the cell in that row containing the strictly compatible item. This function can calculate linear or square weights using the original and augmented datasets of the DenseNet model as shown in Tables 15 and 16.

Comparison of experimental analysis

This study used four datasets: JAFFE, CK+, AffectNet, and the Custom dataset. The model has been tested and trained with validation and loss in which the accuracy and loss have been illustrated. On the datasets mentioned above and validation with loss and accuracy, we show the performance of our proposed DFER model. We briefly discuss our training approach before we further the evaluation process. We have tried to keep the architecture and hyperparameters used for testing and training the datasets. To analyze the weight of the network, we used the Gaussian Random Integer variables, in which the weight was kept at 0.05 with zero convolution. Also utilized Adam's optimizer with the proper rate of 0.005 values. Apart from this, a diverse optimizer was also utilized, including the Adam and lowering the stochastic gradients, which seemed to be more successfully obtained as the weight was decreased up to 0.001 value. For version L2 augmentation, it was used for magnification purposes. At the FER-2013, JAFFE, CK+, and AffectNet, our proposed custom-made dataset and the model achieved the best performance and took 15 days.

On the other hand, our custom dataset has taken only 4 days to accomplish. We have used minor distortions, small rotations, and reversals to improve the data. We use oversampling for model training on classes with fewer images in the

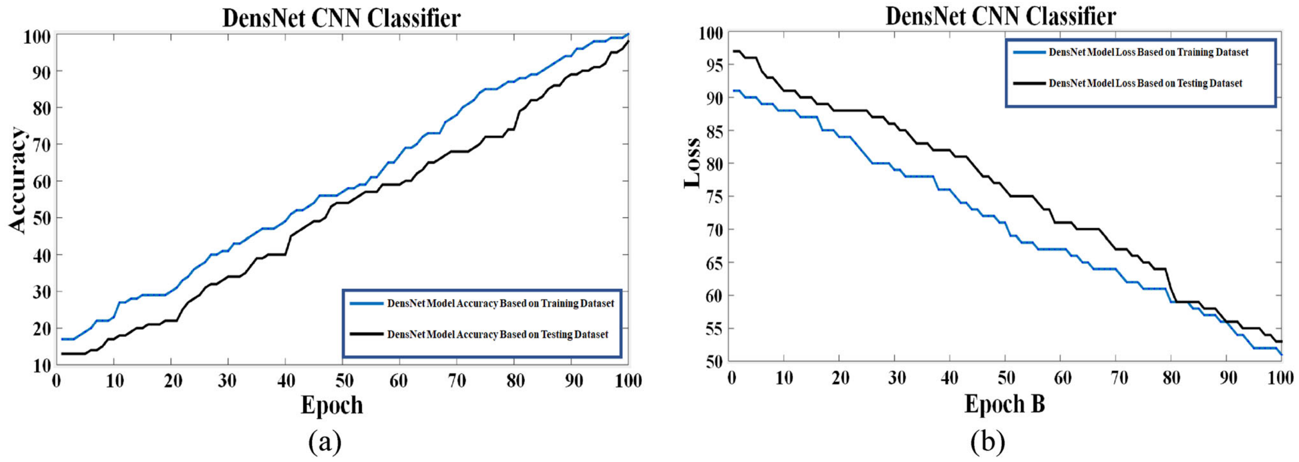


Fig. 28 DeneNet CNN model based on training and validation, **a** accuracy and **b** loss plot for CDD

Table 11 Detailed test accuracy by class of CDD original

Class	Accuracy	Precision	Recall	F-measure
Happy	99.39	96.99	98.25	97.81
Anger	99.58	99.22	99.25	97.65
Fear	96.93	99.99	98.56	98.39
Disgust	99.3	98.37	98.12	98.24
Sad	99.1	98.36	97.01	97.88
Surprise	99.93	99.32	98.53	98.53
Neutral	99.91	99.56	99.37	97.56
Average	99.16	98.83	98.44	98

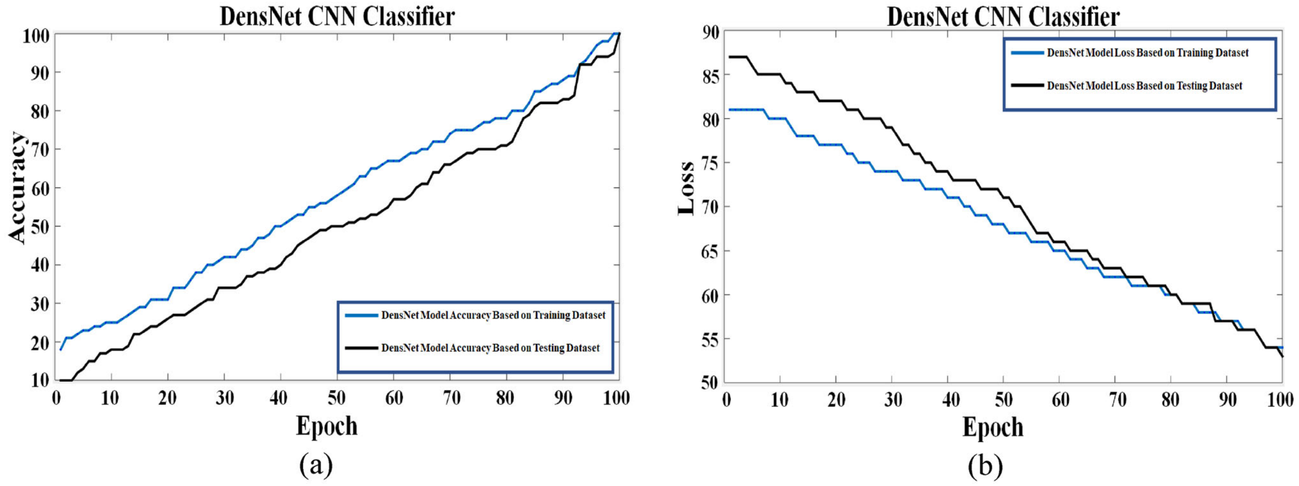


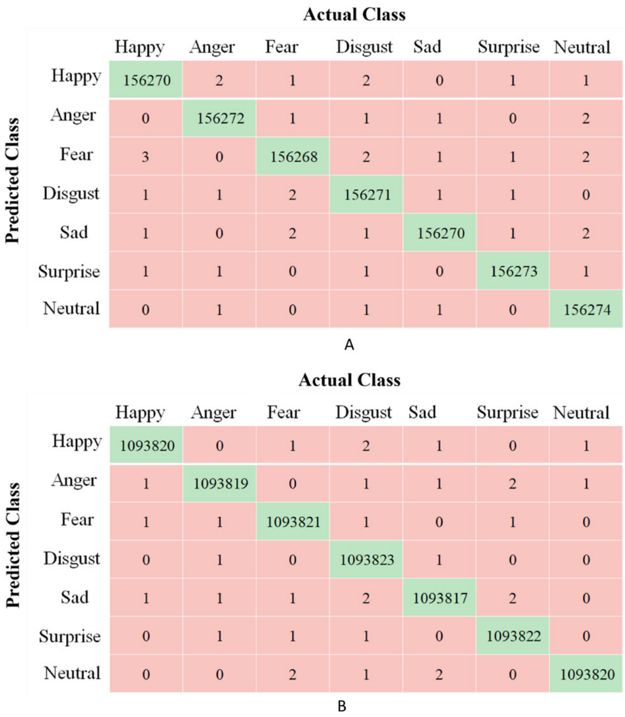
Fig. 29 DeneNet CNN model based on training and validation, **a** accuracy and **b** loss plot for augmented CDD

dataset, which solves the data imbalance problem and leads to model generalization. Classes can have the same order to train huge models.

With the use of FER-2013, the other recognition datasets of DFER are more accessible. Apart from these, in the FER, with the variation in the internal class of the datasets, the additional research challenge is the unbalanced nature of the diverse emotion classes. Emotions such as neutrality and happiness have more examples as compared to others. The proposed model was tested and trained using the 28,709 images as input for the set of training images, which were verified with the validation images of 3500 and tested with the images of 3589.

Table 12 Detailed test accuracy by class of augmented CDD

Class	Accuracy	Precision	Recall	F-measure
Happy	99.99	99.93	99.98	99.89
Anger	99.93	99.98	99.98	99.91
Fear	99.98	99.96	99.99	99.86
Disgust	99.97	99.99	99.97	99.55
Sad	99.98	99.98	99.97	99.59
Surprise	99.99	99.98	99.88	99.36
Neutral	99.96	99.99	99.89	99.91
Average	99.97	99.97	99.95	99.72



Actual Class

Predicted Class	Happy	Anger	Fear	Disgust	Sad	Surprise	Neutral
Happy	156270	2	1	2	0	1	1
Anger	0	156272	1	1	1	0	2
Fear	3	0	156268	2	1	1	2
Disgust	1	1	2	156271	1	1	0
Sad	1	0	2	1	156270	1	2
Surprise	1	1	0	1	0	156273	1
Neutral	0	1	0	1	1	0	156274

A

Actual Class

Predicted Class	Happy	Anger	Fear	Disgust	Sad	Surprise	Neutral
Happy	1093820	0	1	2	1	0	1
Anger	1	1093819	0	1	1	2	1
Fear	1	1	1093821	1	0	1	0
Disgust	0	1	0	1093823	1	0	0
Sad	1	1	1	2	1093817	2	0
Surprise	0	1	1	1	0	1093822	0
Neutral	0	0	2	1	2	0	1093820

B

Fig. 30 Confusion matrix of CDD original (**a**) and **b** augmented

Using the FER-2013 original dataset of the testing sets, there has been an accuracy of 99.01% achieved. While the FER-2013 augmented dataset, the DenseNet model's achieved average accuracy is 99.27% under the simulation environment compared with the benchmark datasets. The year 2013 was determined with the computation of the outcomes for our proposed model, which is illustrated with the current work in Fig. 34 on FER-2013 and the accuracy graph.

We have used 120 images for training the JAFFE dataset, although 23 images were used for validation purposes for the JAFFE dataset. We used 70 for JAFFE dataset testing. The overall accuracy of the JAFFE original dataset is 89.2%. Compared with benchmark datasets, the model for the augmented JAFFE dataset achieves an accuracy of 98.01.01%, as shown in Fig. 35.

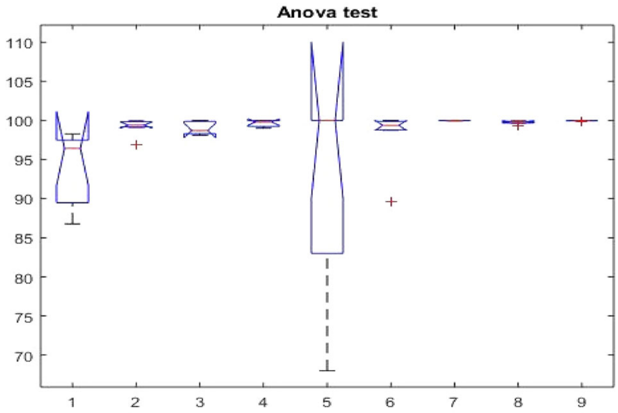


Fig. 31 Anova test of DenseNet model

Figure 36 represents the proposed dataset with benchmark datasets, and all datasets trained by the simulator in a simulated environment and compared with the benchmark dataset are limited to 120 images. For validation, the CK+ dataset and other as used 23 and 70 images are used to test the datasets under the simulation environment. The overall accuracy of the CK+ original dataset is 97.30% for an expanded dataset, while the augmented CK+ dataset obtained 98.53%. The proposed model obtained better accuracy compared with state-of-the-art datasets under the simulation environment, which are given in Fig. 36.

In this scenario, the validation for the AffectNet dataset uses 23 images, and 70 images are used for training the datasets under the simulation environment. The overall accuracy of the AffectNet original dataset is 87.37% for an expanded dataset, while the augmented AffectNet dataset obtained 98.81%. The proposed model obtained better accuracy compared with state-of-the-art datasets under the simulation environment, which are shown in Fig. 37.

Conclusions

In this paper, a new framework based on CNN is proposed to recognize the emotional state of the driver. We

Table 13 Multiple comparison results in a Table 36 × 6 table

Group A	Group B	Lower limit	A-B	Upper limit	*p* value
1	2	− 13.7765394	− 5.505714286	2.765110824	0.451446932
1	3	− 13.64082511	− 5.37	2.90082511	0.485644913
1	4	− 14.22939654	− 5.958571429	2.312253682	0.344571994
1	5	− 6.327967967	1.942857143	10.21368225	0.997513056
1	6	− 12.66511082	− 4.394285714	3.876539396	0.733911489
1	7	− 14.58511082	− 6.314285714	1.956539396	0.271133549
1	8	− 14.34225368	− 6.071428571	2.199396539	0.320155605
1	9	− 14.56511082	− 6.294285714	1.976539396	0.27498209
2	3	− 8.135110824	0.135714286	8.406539396	1
2	4	− 8.723682253	− 0.452857143	7.817967967	0.999999965
2	5	− 0.822253682	7.448571429	15.71939654	0.109112258
2	6	− 7.159396539	1.111428571	9.382253682	0.9999596
2	7	− 9.079396539	− 0.808571429	7.462253682	0.99999656
2	8	− 8.836539396	− 0.565714286	7.705110824	0.999999793
2	9	− 9.059396539	− 0.788571429	7.482253682	0.999997172
3	4	− 8.859396539	− 0.588571429	7.682253682	0.999999717
3	5	− 0.957967967	7.312857143	15.58368225	0.12298599
3	6	− 7.295110824	0.975714286	9.246539396	0.999985169
3	7	− 9.215110824	− 0.944285714	7.326539396	0.99998849
3	8	− 8.972253682	− 0.701428571	7.569396539	0.999998871
3	9	− 9.195110824	− 0.924285714	7.346539396	0.99999025
4	5	− 0.369396539	7.901428571	16.17225368	0.071797481
4	6	− 6.706539396	1.564285714	9.835110824	0.999474046
4	7	− 8.626539396	− 0.355714286	7.915110824	0.999999995
4	8	− 8.383682253	− 0.112857143	8.157967967	1
4	9	− 8.606539396	− 0.335714286	7.935110824	0.999999997
5	6	− 14.60796797	− 6.337142857	1.933682253	0.26677755
5	7	− 16.52796797	− 8.257142857	0.013682253	0.050690287
5	8	− 16.28511082	− 8.014285714	0.256539396	0.064404106
5	9	− 16.50796797	− 8.237142857	0.033682253	0.051714333
6	7	− 10.19082511	− 1.92	6.35082511	0.997710696
6	8	− 9.947967967	− 1.677142857	6.593682253	0.999126535
6	9	− 10.17082511	− 1.9	6.37082511	0.997872898
7	8	− 8.027967967	0.242857143	8.513682253	1
7	9	− 8.25082511	0.02	8.29082511	1
8	9	− 8.493682253	− 0.222857143	8.047967967	1

Table 14 Anova table of DenseNet model

Source	SS	*Df*	MS	*F*	Prob > *F*
Columns	517.54	8	64.6925	2.82	0.0109
Error	1238.66	54	22.9383		
Total	1756.2	62			

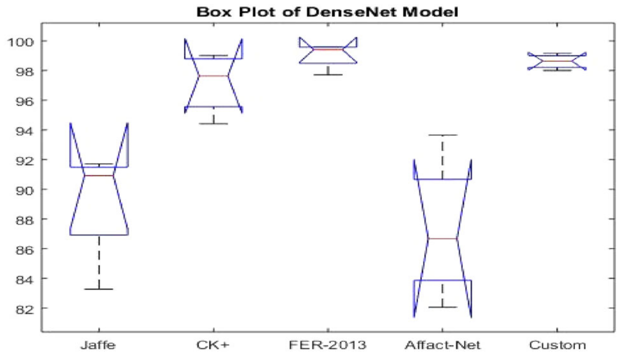


Fig. 32 DenseNet Model using original datasets

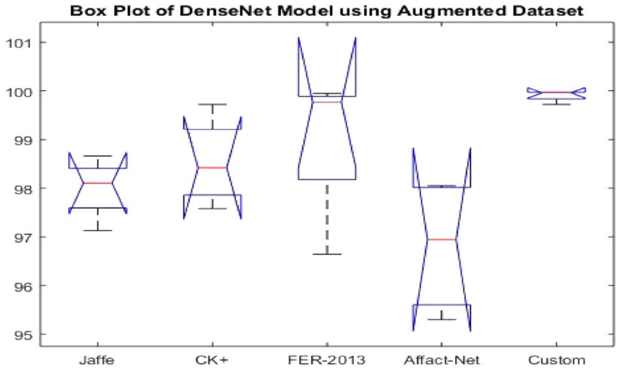


Fig. 33 DenseNet model using augmented datasets

Table 15 Unweighted Cohen's Kappa for DenseNet using original datasets

Cohen's kappa	0.0004
Kappa CI (alpha = 0.0500)	− 0.0220 0.0228
Kappa error	0.0114
Agreement due to true concordance (po-pe)	0.0003
k observed as a proportion of the maximum possible	0.0005
Random agreement (pe)	0.1973
Maximum possible kappa, given the observed marginal frequencies	0.7379
Residual not random agreement (1-pe)	0.8027
Observed agreement (po)	0.1977
Slight agreement	
z (k/kappa error)	0.0354 $p = 0.9718$

Table 16 Unweighted Cohen's kappa for DenseNet using augmented datasets

Cohen's kappa	− 0.0003
Kappa error	0.0112
Random agreement (pe)	0.1996
Maximum possible kappa, given the observed marginal frequencies	0.7483
Residual not random agreement (1-pe)	0.8004
Observed agreement (po)	0.1994
Agreement due to true concordance	− 0.0002
(po-pe) Kappa CI (alpha = 0.0500)	− 0.0223 0.0217
k observed as a proportion of the maximum possible	− 0.0004
Poor agreement	
z (k/kappa error)	− 0.0273 $p = 0.9782$

believe that FE can be detected by focusing on specific facial regions. We have also conducted experimental research, utilizing four different expression datasets and a custom-made dataset for DFER, which yielded promising results. In addition, we employed a visualization technique to highlight the most crucial areas of face images to recognize various drivers' facial emotions. The model identifies the driver's emotions using images of the driver's face with the help of a deep learning algorithm. According to the proposed DFER model, a DFE state can be identified without additional effort from the driver. A custom CNN block replaces the Improved Faster R-CNN features learning block to improve the accuracy and efficiency of face detection. Transfer learning is implemented in DenseNet CNN's model by substituting custom driver emotion datasets for ImageNet data. The CNN model consists of 201 layers and is used to recognize facial expressions. The CDD contains seven basic driver emotions, and the proposed face detection and DFER models are evaluated using the benchmark dataset. The effectiveness of the DenseNet model has been evaluated by achieving high accuracy. Multiple facial expression recognition datasets were utilized to evaluate the proposed model, including JAFFE, CK+, FER-2013, AffectNet, and custom-made datasets. The proposed model outperformed several advanced facial expression recognition models using a benchmark dataset. The DenseNet model is efficient and precise and can be implemented by various hardware devices to recognize the driver's emotions and improve the performance of automatic driver assistance systems.

Putting aside what we have successfully achieved, several useful extensions can be addressed for further improvements. Without considering the influence of head pose variations, only the frontal faces of drivers are taken for training and implementation purposes. Therefore, further faces from several views can be considered from the images or videos which may help to improve the recognition accuracy. The deep learning techniques lack sufficient data to be the most effective they can. Therefore, it may be useful to pre-train a deep CNN on many other datasets before applying a fine-tuning process. A hybrid method can be developed in the future by

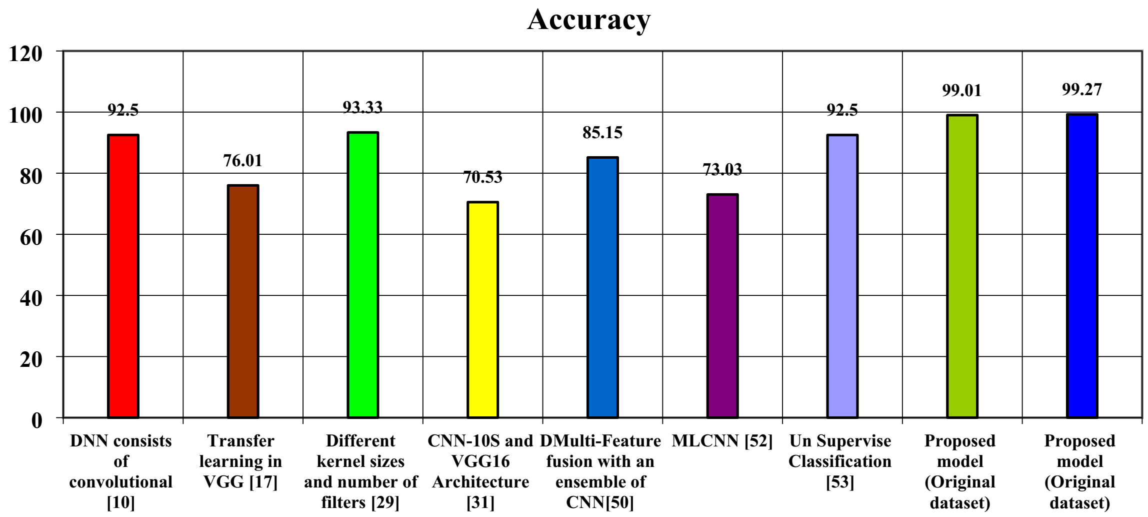


Fig. 34 Classification accuracy of the proposed FER-2013 dataset and comparison

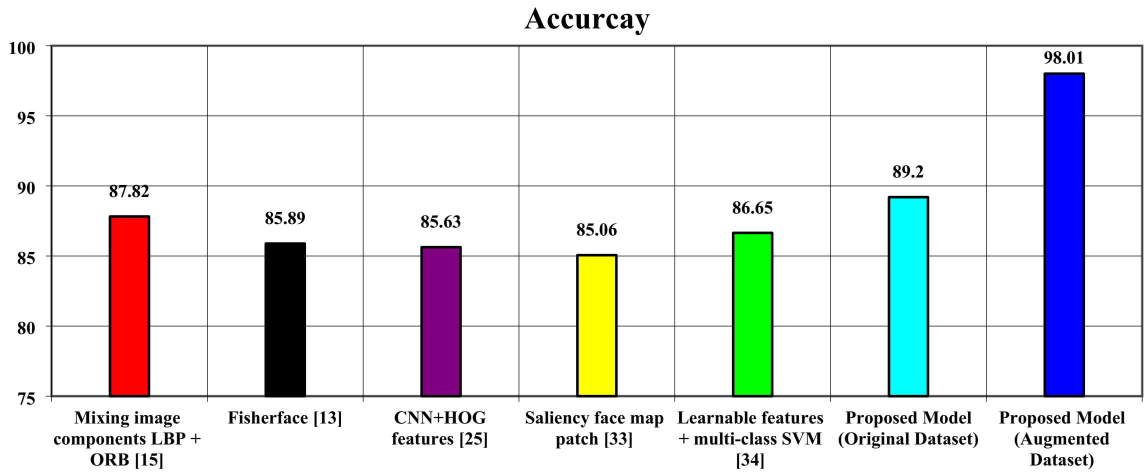


Fig. 35 Classification accuracy of the original and augmented JAFFE dataset

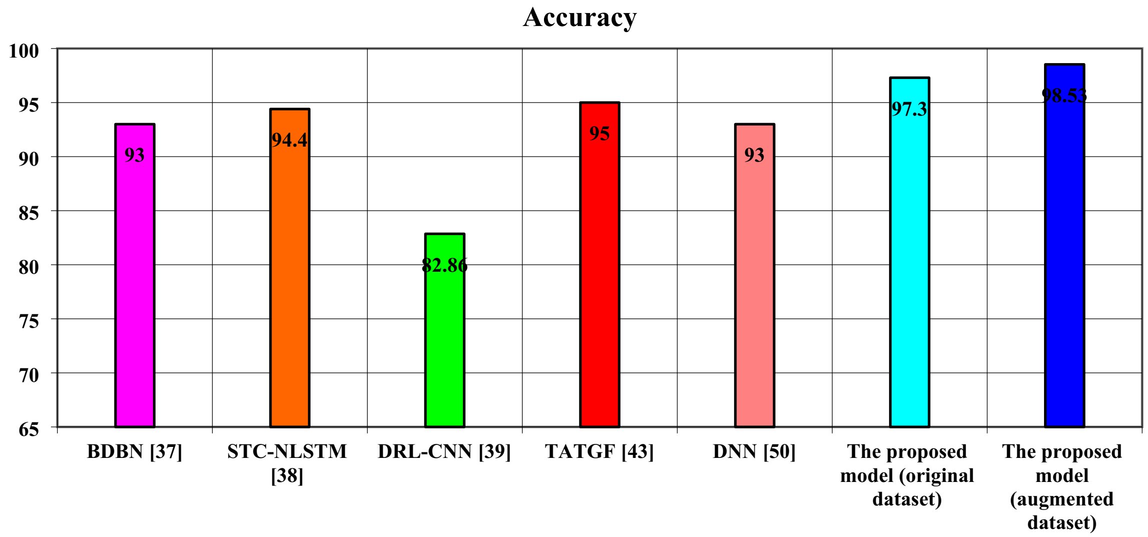


Fig. 36 The proposed CK+ dataset original and augmented classification accuracy, and comparison with other datasets

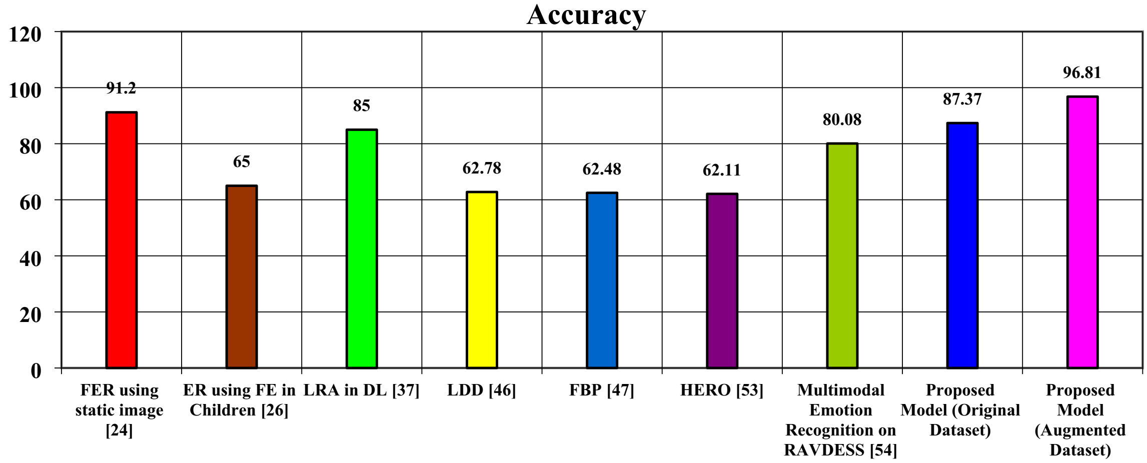


Fig. 37 Classification accuracy of the proposed original and augmented AffectNet dataset

combining geometric features and appearance-based features to improve the performance of the DER.

Author contributions KZ, TH and SZ conceptualized the study and obtained the resources. KZ and SZ performed data concoction and administrated the project. KZ, TH, BS and SZ performed the formal analysis. KZ, BS and SZ prepared methodology, investigated the study, and wrote the original draft preparation. SZ was responsible for funding acquisition. KZ. FA and SZ validated the study. KZ, SZ, SMS, AH, USK, MS, and RU reviewed and edited the manuscript. All the authors have read and agreed to the published version of the manuscript.

Funding This research was funded by the National Key Research and Development Program, grant numbers 2018YFB1600202 and 2021YFB1600205; National Natural Science Foundation of China, grant number 52178407 and 62172366. This research work is supported by RIF activity code 23009 of Zayed University, UAE.

Data availability The dataset used in this research is available on the following web link: https://github.com/tariqaup/Driver-Emotion-Recognition-System; JAFFE dataset to visit: https://zenodo.org/records/3451524; CK+ dataset to visit: http://vasc.ri.cmu.edu/idb/html/face/facial_expression; FER dataset to visit: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data.

Declarations

Conflict of interest All the authors declare no conflict of interest.

References

1. Pena-Garijo J, Lacruz M, Masanet MJ, Palop-Grau A, Plaza R, Hernandez-Merino A, Valllina O (2023) Specific facial emotion recognition deficits across the course of psychosis: a comparison of individuals with low-risk, high-risk, first-episode psychosis and multi-episode schizophrenia-spectrum disorders. Psychiatry Res 320:115029
2. Meng Q, Hu X, Kang J, Wu Y (2020) On the effectiveness of facial expression recognition for evaluation of urban sound perception. Sci Total Environ 710:135484
3. Hussain T, Yang B, Rahman HU, Iqbal A, Ali F (2022) Improving source location privacy in social internet of things using a hybrid phantom routing technique. Comput Secur 123:102917
4. Akter T, Ali MH, Khan MI, Satu MS, Uddin MJ, Alyami SA, Moni MA (2021) Improved transfer-learning-based facial recognition framework to detect autistic children at an early stage. Brain Sci 11(6):734
5. Rahul M, Tiwari N, Shukla R, Tyagi D, Yadav V (2022) A new hybrid approach for efficient emotion recognition using deep learning. Int J Electr Electron Res (IJEER) 10(1):18–22
6. Leo M, Carcagnì P, Distante C, Spagnolo P, Mazzeo PL, Rosato AC, Lecciso F (2018) Computational assessment of facial expression production in ASD children. Sensors 18(11):3993
7. Sun S, Ge C (2014) A new method of 3D facial expression animation. J Appl Math 2014:1–6
8. Saste ST, Jagdale SM (2017) Emotion recognition from speech using MFCC and DWT for security system. In: 2017 international conference of electronics, communication and aerospace technology (ICECA) (Vol. 1, pp 701–704). IEEE
9. Houssein EH, Hammad A, Ali AA (2022) Human emotion recognition from EEG-based brain–computer interface using machine learning: a comprehensive review. Neural Comput Appl 34(15):12527–12557
10. Mollahosseini A, Chan D, Mahoor MH (2016) Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 2016, pp. 1–10, https://doi.org/10.1109/WACV.2016.7477450
11. Liu P, Han S, Meng Z, Tong Y (2014) Facial expression recognition via a boosted deep belief network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Columbus, OH, USA, pp 1805–1812

12. Kumar P, Raman B (2022) A BERT based dual-channel explainable text emotion recognition system. Neural Netw 150:392–407
13. Han K, Yu D, Tashev I (2014) Speech emotion recognition using deep neural network and extreme learning machine. In: Interspeech 2014
14. Hussain A, Ahmad M, Hussain T, Ullah I (2022) Efficient content based video retrieval system by applying AlexNet on key frames. ADCAIJ Adv Distrib Comput Artif Intell J 11(2):207–235
15. Bouzidi M, Barkat S, Krama A, Abu-Rub H (2022) Generalized predictive direct power control with constant switching frequency for multilevel four-leg grid connected converter. IEEE Trans Power Electron 37(6):6625–6636
16. Khorrami P, Paine T, Huang T (2015) Do deep neural networks learn facial action units when doing expression recognition. In: Proceedings of the IEEE international conference on computer vision workshops, Santiago, Chile, pp 19–27
17. Tzirakis P, Trigeorgis G, Nicolaou MA, Schuller BW, Zafeiriou S (2017) End-to-end multimodal emotion recognition using deep neural networks. IEEE J Select Top Signal Process 11(8):1301–1309
18. Zhang W et al. (2022) Transformer-based Multimodal Information Fusion for Facial Expression Analysis. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 2022, pp 2427–2436, https://doi.org/10.1109/CVPRW56347.2022.00271
19. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13. Springer International Publishing, pp 818–833
20. Bolioli A, Bosca A, Damiano R, Lieto A, Striani M (2022) A complementary account to emotion extraction and classification in cultural heritage based on the Plutchik's theory. In: Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, New York, NY, United States, pp 374–382. https://doi.org/10.1145/3511047.3537659
21. Correia-Caeiro C, Burrows A, Wilson DA, Abdelrahman A, Miyabe-Nishiwaki T (2022) CalliFACS: the common marmoset facial action coding system. PLoS ONE 17(5):e0266442
22. Huo Y, Zhang L (2022) OCFER-Net: recognizing facial expression in online learning system. In: Proceedings of the 2022 International Conference on Advanced Visual Interfaces (AVI 2022). Association for Computing Machinery, New York, NY, USA, Article 45, 1–3. https://doi.org/10.1145/3531073.3534470
23. Zhang Y, Zhang Q, Yang J (2022) Application of an artificial intelligence system recognition based on the deep neural network algorithm. Comput Intell Neurosci 2022:4623188. https://doi.org/10.1155/2022/4623188
24. Ge H, Zhu Z, Dai Y, Wang B, Wu X (2022) Facial expression recognition based on deep learning. Comput Methods Programs Biomed 215:106621
25. Karras C, Karras A, Sioutas S (2022) Pattern recognition and event detection on IoT data-streams. arXiv preprint. https://arXiv.org/2203.01114https://doi.org/10.48550/arXiv.2203.01114
26. Chen J, Chen Z, Chi Z, Fu H (2014) Facial expression recognition based on facial components detection and hog features. In: Scientific Cooperations International Workshops on Electrical and Computer Engineering Subfields 22–23 August 2014, Koc University, Istanbul/Turkey, pp 884–888
27. Barsoum E, Zhang C, Canton Ferrer C, Zhang Z (2016) Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI '16). Association for Computing Machinery, New York, NY, USA, 279–283. https://doi.org/10.1145/2993148.2993165
28. Han S, Meng Z, Khan AS, Tong Y (2016) Incremental boosting convolutional neural network for facial action unit recognition. In: Advances in neural information processing systems, 29
29. Meng Z, Liu P, Cai J, Han S, Tong Y (2017) Identity-aware convolutional neural network for facial expression recognition. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 2017, pp. 558–565, https://doi.org/10.1109/FG.2017.140
30. Fernandez P, Pena F, Ren T, Cunha A (2019) FERAtt: facial expression recognition with attention net. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 2019, pp. 837–846. https://doi.org/10.1109/CVPRW.2019.00112
31. Wang K, Peng X, Yang Y, Lu S, Qiao Y (2020) Suppressing uncertainties for large-scale facial expression recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp 6896–6905, https://doi.org/10.1109/CVPR42600.2020.00693
32. Wang K, Peng X, Yang J, Meng D, Qiao Y (2020) Region attention networks for pose and occlusion robust facial expression recognition. IEEE Trans Image Process 29:4057–4069
33. Gupta A, Arunachalam S, Balakrishnan R (2020) Deep self-attention network for facial emotion recognition. Procedia Comput Sci 171:1527–1534
34. Gan Y, Chen J, Yang Z, Xu L (2020) Multiple attention network for facial expression recognition. IEEE Access 8:7383–7393
35. Li S, Deng W (2022) Deep facial expression recognition: a survey. IEEE Trans Affect Comput 13(3):1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446
36. Jaderberg M, Simonyan K, Zisserman A (2015) Spatial transformer networks. Adv Neural Inform Process Syst 28
37. Wang X, Wang X, Ni Y (2018) Unsupervised domain adaptation for facial expression recognition using generative adversarial networks. Comput Intell Neurosci 2018:7208794. https://doi.org/10.1155/2018/7208794
38. Khalid M, Keming M, Hussain T (2021) Design and implementation of clothing fashion style recommendation system using deep learning. Rom J Inform Technol Autom Control 31(4):123–136
39. Georgescu MI, Ionescu RT, Popescu M (2019) Local learning with deep and handcrafted features for facial expression recognition. IEEE Access 7:64827–64836
40. Giannopoulos P, Perikos I, Hatzilygeroudis I (2018) Deep learning approaches for facial emotion recognition: a case study on FER-2013. In: Hatzilygeroudis I, Palade V (eds) Advances in hybridization of intelligent methods smart innovation, systems and technologies, vol 85. Springer, Cham. https://doi.org/10.1007/978-3-319-66790-4_1
41. Minaee S, Minaei M, Abdolrashidi A (2021) Deep-emotion: facial expression recognition using attentional convolutional network. Sensors 21(9):3046
42. Kollias D, Zafeiriou S (2019) Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. arXiv preprint. https://arXiv.org/1910.04855https://doi.org/10.48550/arXiv.1910.04855
43. Niu B, Gao Z, Guo B (2021) Facial expression recognition with LBP and ORB features. Comput Intell Neurosci 2021:8828245. https://doi.org/10.1155/2021/8828245
44. Irfanullah, Hussain T, Iqbal A et al (2022) Real time violence detection in surveillance videos using convolutional neural networks. Multimed Tools Appl 81:38151–38173. https://doi.org/10.1007/s11042-022-13169-4
45. Wang H, Wei S, Fang B (2020) Facial expression recognition using iterative fusion of MO-HOG and deep features. J Supercomput 76(5):3211–3221
46. Bhattacharya S (2022) A survey on: facial expression recognition using various deep learning techniques. Advanced computational

paradigms and hybrid intelligent computing. Springer, Singapore, pp 619–631

47. Shima Y, Omori Y (2018) Image augmentation for classifying facial expression images by using deep neural network pre-trained with object image database. In: Proceedings of the 3rd International Conference on Robotics, Control and Automation (ICRCA'18). Association for Computing Machinery, New York, NY, USA, 140–146. https://doi.org/10.1145/3265639.3265664
48. Wang SH, Zhang YD (2020) DenseNet-201-based deep neural network with composite learning factor and precomputation for multiple sclerosis classification. ACM Trans Multimed Comput Commun Appl (TOMM) 16(2s):1–19
49. Zaman K, Sun Z, Shah SM, Shoaib M, Pei L, Hussain A (2022) Driver Emotions recognition based on improved faster R-CNN and neural architectural search network. Symmetry 14(4):687
50. Sohail A, Khan A, Nisar H, Tabassum S, Zameer A (2021) Mitotic nuclei analysis in breast cancer histopathology images using deep ensemble classifier. Med Image Anal 72:102121
51. Hong S, Wu M, Zhou Y, Wang Q, Shang J, Li H, Xie J (2017) ENCASE: an ENsemble ClASsifiEr for ECG classification using expert features and deep neural networks. In: 2017 Computing in cardiology (cinc). IEEE, pp. 1–4
52. Gu X, Angelov PP, Zhang C, Atkinson PM (2018) A massively parallel deep rule-based ensemble classifier for remote sensing scenes. IEEE Geosci Remote Sens Lett 15(3):345–349
53. Potamias RA, Siolas G, Stafylopatis A (2019) A robust deep ensemble classifier for figurative language detection. In: Macintyre J, Iliadis L, Maglogiannis I, Jayne C (eds) Engineering applications of neural networks. EANN 2019. Communications in Computer and Information Science, vol 1000. Springer, Cham. https://doi.org/10.1007/978-3-030-20257-6_14
54. Shakeel PM, Burhanuddin MA, Desa MI (2022) Automatic lung cancer detection from CT image using improved deep neural network and ensemble classifier. Neural Comput Appl 34:9579–9592. https://doi.org/10.1007/s00521-020-04842-6
55. Shah SM, Sun Z, Zaman K, Hussain A, Shoaib M, Pei L (2022) A driver gaze estimation method based on deep learning. Sensors 22(10):3959
56. Ullah R, Gani A, Shiraz M, Yousufzai IK, Zaman K (2022) Auction mechanism-based sectored fractional frequency reuse for irregular geometry multicellular networks. Electronics 11(15):2281
57. Malakar S, Ghosh M, Bhowmik S, Sarkar R, Nasipuri M (2020) A GA based hierarchical feature selection approach for handwritten word recognition. Neural Comput Appl 32(7):2533–2552
58. Bacanin N, Stoean R, Zivkovic M, Petrovic A, Rashid TA, Bezdan T (2021) Performance of a novel chaotic firefly algorithm with enhanced exploration for tackling global optimization problems: application for dropout regularization. Mathematics 9(21):2705
59. Lyons MJ, Akamatsu S, Kamachi M, Gyoba J (1998) Coding facial expressions with Gabor wavelets. In: 3rd IEEE International Conference on Automatic Face and Gesture Recognition, pp 200–205. https://doi.org/10.1109/AFGR.1998.670949. Open access content available at: https://zenodo.org/record/3430156