symmetry

MDPI

Article

Driver Emotions Recognition Based on Improved Faster R-CNN and Neural Architectural Search Network

Khalid Zaman [1], Zhaoyun Sun [1,*], Sayyed Mudassar Shah [1], Muhammad Shoaib [2], Lili Pei [1] and Altaf Hussain [3]

[1] Information Engineering School, Chang'an University, Xi'an 710061, China; khalidzaman@chd.edu.cn (K.Z.); mudassarshah@chd.edu.cn (S.M.S.); peilili@chd.edu.cn (L.P.)
[2] Department of Computer Science and IT, CECOS University, Peshawar 25000, Pakistan; mshoaib@cecos.edu.pk
[3] Institute of Computer Science and IT, The University of Agriculture, Peshawar 25000, Pakistan; altafscholar@aup.edu.pk
* Correspondence: chysun@chd.edu.cn; Tel.: +86-13572190029

Abstract: It is critical for intelligent vehicles to be capable of monitoring the health and well-being of the drivers they transport on a continuous basis. This is especially true in the case of autonomous vehicles. To address the issue, an automatic system is developed for driver's real emotion recognizer (DRER) using deep learning. The emotional values of drivers in indoor vehicles are symmetrically mapped to image design in order to investigate the characteristics of abstract expressions, expression design principles, and an experimental evaluation is conducted based on existing research on the design of driver facial expressions for intelligent products. By substituting a custom-created CNN features learning block with the base 11 layers CNN model in this paper for the development of an improved faster R-CNN face detector that detects the driver's face at a high frame per second (FPS). Transfer learning is performed in the NasNet large CNN model in order to recognize the driver's various emotions. Additionally, a custom driver emotion recognition image dataset is being developed as part of this research task. The proposed model, which is a combination of an improved faster R-CNN and transfer learning in NasNet-Large CNN architecture for DER based on facial images, enables greater accuracy than previously possible for DER based on facial images. The proposed model outperforms some recently updated state-of-the-art techniques in terms of accuracy. The proposed model achieved the following accuracy on various benchmark datasets: JAFFE 98.48%, CK+ 99.73%, FER-2013 99.95%, AffectNet 95.28%, and 99.15% on a custom-developed dataset.

Keywords: driver emotions recognition; computer vision; facial expression recognition; facial image symmetry; improved faster R-CNN; neural architecture search network

Citation: Zaman, K.; Sun, Z.; Shah, S.M.; Shoaib, M.; Pei, L.; Hussain, A. Driver Emotions Recognition Based on Improved Faster R-CNN and Neural Architectural Search Network. *Symmetry* **2022**, *14*, 687. https://doi.org/10.3390/sym14040687

Academic Editors: Gianluca Vinti and Sergei D. Odintsov

Received: 7 February 2022
Accepted: 22 March 2022
Published: 26 March 2022

1. Introduction

Drivers' emotional states can impact their driving ability while driving a vehicle [1,2]. Due to the increasing sophistication of vehicles, recognizing the emotions of their drivers becomes more and more important. To ensure a more secure and pleasant ride, good infotainment can precisely detect the driver's emotional state before making adjustments to the vehicle's dynamics. In intelligent cars, it is critical to recognize the emotions of the driver because the vehicle can make decisions about what to do in certain situations based on the driver's psychological state (for example driving modes, mood-altering songs, and autonomous driving). Facial expressions (FEs) are considered necessary in human-machine interfaces because they aid in expressing human emotions and feelings, which is essential in developing artificial intelligence. A new research area called facial expression recognition (FER) has been established. Recent years have seen significant advancements in deep learning-based image recognition techniques [3–8], and deep learning is becoming increasingly popular for FER. Although a person's facial expressions often accurately reflect their genuine emotions, various factors can influence how accurately they do so.

Drivers, in particular, demonstrate a more pronounced manifestation of this trait. Pretend you are in the following situation: The presence of a worried expression on a driver's face while driving may lead a reasonable observer to conclude that if only the driver's facial expressions are taken into account while making assumptions, the car driver is presently in an unhappy state. The opposite is true if the only muscles in a driver's face react to sunlight stimuli, which does not indicate that the driver is experiencing any discomfort. As a result, the driver's emotions do not always show up in their facial expressions when they are driving. Therefore, our goal is to identify the genuine emotions of a driving partner, even when these emotions cannot be fully expressed through facial expressions while driving.

Similarly, microfacial expressions, which investigates subtle changes in facial expressions over a brief period, are closely related to those described above. When real emotions are suppressed, whether intentionally or unintentionally, it is common for such minor changes to occur due to the suppression. Researchers have developed promising methods for detecting hidden emotions based on several studies focusing on facial micro-expressions [9–12]. Using micro-expressions in conjunction with deep learning-based algorithms, drivers' genuine emotions can be determined. However, the shortage of samples and an unbalanced division of samples are the key barriers to widespread use. In the end, we are looking for drivers who are experiencing genuine emotions rather than drivers who have suppressed or hidden their emotions from view [13].

Furthermore, most studies [14–23] use physiological signals to identify human emotions, which is a significant advance. In clinical practice, the electroencephalogram (EEG), the electrocardiogram (ECG), the photoplethysmography (PPG), and the electrical skin activity (ESA) are the physiological signals that are most often used (EDA). A few studies have found that a combination of both facial expressions and physiological signals may be used to accurately identify [24,25] a wide range of emotions [26,27]. This research is based on deep learning algorithms, which are gaining popularity at an alarming rate. A deep learning-based DRER that uses sensor fusion of driver FER and physiological data to recognize the actual emotional state of the driver while driving is proposed in this paper based on these trends. The DRER is intended to detect the actual emotional state of the driver while the vehicle is in motion.

The rest of the paper is structured as follows: Section two discusses the previous sentiment recognition research and its application today. The third section discusses the DRER, a technique for detecting the driver's genuine emotions while driving. It is proposed that the proposed DRER make use of the facial expressions of driver and EDA data collected while driving. Section 3 contains a more detailed description of the proposed model. The fifth page contains a detailed experimental design and a simulation with a human in the loop. Section 4 compares and analyses the findings from experimental results. Section 5 concludes the proposed work with future work recommendations.

2. Literature Review

This section summarizes research in the field of recognition of emotion. This chapter discusses FER, emotion recognition via biophysiological signals, and synchronous emotion recognition (SFER). Additionally, we implement a dataset and a simulation with a human in the loop to ensure the security of emotion recognition data.

2.1. Facial Expression Recognition (FER)

The majority of prior research has generated statistically significant results by utilizing CNNs and other networks such as support vector machines (SVMs) and recurrent neural networks (RNNs) and preprocessing the data. The Emotional Recognition in Wild Challenges compared their results by using a variety of techniques of deep learning, including convolutional neural networks, deep belief networks, relational autoencoders, and shallow neural networks. Additionally, they asserted that the optimal model could extract sentiment class probabilities and train SVM hyperparameters using CNNs, rather than one that

relied solely on machine learning [1]. By contrast, FER is unique because it is founded on the definition of emotions rather than on any other concept. Emotions are classified into discrete states of mind and continuous states of mind. Categorizing discrete emotions such as rage, disgust, happiness, and neutrality is accomplished at the category level. A face detector and similarity transform were used to preprocess in the Wild (AFEW) dataset of Acted Facial Expressions, and a CNN filter was used to remove faceless frames in this case. The data were preprocessed before being fed into a CNN (VGG16) and a 3D convolutional model. While the CNN extracts features and maps the input to fixed-length vectors via an RNN/LSTM encoder, the 3D convolutional model encodes motion using video shape and motion [2]. The authors of [3] developed a model specifically for drivers to determine whether or not they are in a stressful state. Stressful situations were defined as those that elicited feelings of rage or disgust. After landmark extraction, they used SIFT descriptors to train the SVM on PCA results. They then applied the features to additional publicly accessible data. They divided the indoor condition (driver's frontal expressions) and the vehicle condition into two groups to collect experimental data for evaluation (expressions from the dashboard). The same is true for driver studies, which typically train models using publicly available data, evaluate those using experimental data, and then define emotions based on the study's objectives. Another way to define emotions is to use a continuous definition. Russel's V-A (valence-arousal) model has been used extensively in previous research on continuous emotion. The term "valence" refers to the strength of an emotional attraction, either negative or positive. Arousal is a term that refers to the physiological activity of various nerves in response to stimulation; it increases in direct proportion to the intensity of the emotion. Given that VA can take on values ranging from -1 to 1, VA regression uses a continuous definition. The FATAUVA-net model was proposed in a study on VA regression and is composed of four layers: core, attribute, action unit (AU), and VA. It is based on the MCNN and includes the core, the attribute, the action unit (AU), and the VA layers. MTCNN detects the face in the core and attribute layers of the CelebA dataset and trains the CNN on the detected regions (for example, the face and eyes). The AU layer utilizes the affect-in-the-wild (AFF-wild) dataset to extract facial parts from the attribute layer, and the V-A layer utilizes the V-A dataset to estimate the V-A [4]. Numerous studies have forecasted the future use of V-A regression and emotion classification. The authors of [5] created the dataset by annotating AU and labeling it with the seven most similar basic emotions. They used GAN to provide semi-supervised guidance to the robot after preprocessing with the FFLD2 face detector. A CNN (AFF-WildNet) was used to extract features, which was then fed into an RNN to generate nine outputs. They are based on V-A estimation in two cases and SoftMax's prediction of seven fundamental emotions in the other four. Because V-A is not intuitive in terms of meaning, some studies have used it as a transitional extract for categorizing categorical emotions. The authors of [6] created an automated system that analyses car drivers' facial expressions based on visual acuity (V-A). A pre-trained YOLO V2 was used to locate diver facial videos in datasets such as AFEW-VA and motor trend magazines and extract features from the videos using CNN (ResNet). SVM was used to extract features and then fed into an RNN (LSTM) as input vectors. The resulting V-A was used to predict six different emotions. In previous research on FER, two parallel CNNs were used, with that of the extracted features being combined prior to recognition. Authors in [7] analyzed various types of preprocessed data using both 3D and 2D convolutional models. SoftMax combined and predicted the properties of each generated model. Authors in [8] developed an algorithm that utilizes pooling to separate two network routes regardless of their presence or absence. The intermediate feature extraction stage combines the results of each route before proceeding to the next layer.

2.2. Bio-Physiological Signals

The authors of [9] demonstrated that performance improves as the data type grows larger, regardless of the recognition method used. Numerous studies on the use of biophysiological signals for emotion recognition have been conducted using the sensors mentioned

previously. The authors of [10] used EDA to process driver emotions in three distinct ways in a virtual driving simulation (neutral, stress, and anger). Support vector machines were used to classify the data, resulting in a class II classification. They accurately predicted emotions in 85% of neutral stress and anger cases and about 70% of stress and anger cases. This means that while large emotional segments can be classified, it is more challenging to classify similar segments collectively. The authors of [11] published a comprehensive sensor's review and methods for recognition of human emotion in their journal. According to the authors' paper, EEG analysis, which is typically performed over a range of five frequencies to determine the valence and arousal's average level or detection efficiency of the subject's brain activity, is a fundamental technique. Research into developing new methods for extracting information from EEG data has recently centered on deep learning techniques. Apart from that, scientists have concentrated their efforts on quantifying and evaluating the QRS wave amplitudes and durations. Emotions can be assessed by identifying P or R peaks in the QRS and examining other parameters. Because EDA contains valuable information about the amplitude and frequency of the EDA signal and the decision process, it can be used for emotion recognition and automatic decision detection. By incorporating machine learning algorithms into the system, it is possible to improve emotion recognition accuracy and identify specific emotions associated with various arousal levels. Before defining the peak and generating heart rate variability, PPG signals are filtered with a high-pass filter to remove noise (HRV). Sensors are frequently used in conjunction to compensate for one another's shortcomings. The authors of [12] attempted to recognize emotions while viewing tactilely enhanced multimedia. On a nine-point SAM scale, subjects were asked to rate four different video clips to determine their overall impression of the video. While watching the video clips, subjects recorded their physiological signals and the researchers extracted various features from these signals. After the extracted features were applied to the extracted features, a K-nearest neighbor classifier was used to classify the emotions. The study discovered that PPG-based features had the highest classification accuracy (78.77%) and that combining EEG, EDA, and PPG features increased classification accuracy (79.76%). The use of EEG in research relating to emotion recognition and the brain's response to various stimuli has grown in popularity in recent years, owing to advancements in EEG sensor measurement technology and deep learning. Biophysiological signals are often used as data input in studies involving emotion recognition. Sensors can be used in various ways to collect biophysiological signals, and each signal produced by a sensor has its own unique set of characteristics. The authors of [13] proposed a subject-independent emotion recognition algorithm based on dynamic empirical convolutional neural networks using EEG signals to assess the average level of potency and arousal (DECNN). The transient characteristics of the Shanghai Jiao Tong University Emotional EEG dataset (SEED) were determined to be preserved by using the empirical mode decomposition (EMD) algorithm to filter out the EEG signals in the EEG signal's frequency bands. The dynamic differential entropy (DDE) algorithm was used to extract the properties of EEG signals. Thus, they are able to reflect the temporal and frequency characteristics of an emotional state. A CNN model was created to distinguish between good and negative emotions, and the researchers published their findings. They also revealed that time-frequency features could be represented as two-dimensional matrices, whereas local correlation features could be represented as images. They achieved a 97.56% accuracy rate with these methods. The authors of [14] suggested a deep confidence conditional random field framework with neuroglial chain (DBN-GC) and conditional random field to capture long-term dependencies, contextual information, and correlation information between distinct EEG channels (CRF). The multichannel EEG signals from the DEAP, AMIGOS, and SEED datasets are segmented according to specific time windows, and raw feature vectors are extracted from these segments. Each raw feature vector is fed into a parallel DBN-GC that extracts the high-level representation of the feature vector from the data. The next step is to feed CRF with high-level feature sequences that contain information about the correlation between EEG channels. A sentiment label generator (CRF) generates predicted sentiment label

sequences, and sentiment states can be determined using an automated sentiment state determination layer based on the K-nearest neighbor algorithm. They achieved an average accuracy of 76.58% when using the leave-one-out cross-validation method. As a result, despite the benefits of EEG, some studies have investigated the use of more miniature demanding sensors. This is because acquiring an EEG signal requires the use of specialized equipment. The authors of [15] developed a simple and convenient method for daily emotion monitoring based on a multimodal wearable biosensor network. The purpose of measurement nodes is to collect and transmit signals from multimodal wearable biosensors to sink nodes. It was discovered that the fuzzy rough nearest neighbor (FRNN) algorithm could classify various emotions using a fuzzy threshold that took EEG concentration into account. This method achieves 65.6% accuracy, which is excellent for wearable technology because it minimizes sample classification ranges and interference from noisy samples. The authors of [16] attempted to use EDA as an input feature in another paper that used EDA to identify human emotions unrelated to the topic at hand. An algorithmic CNN model was devised by the researchers, and it predicted four distinct emotional states: high-valence, high-arousal (HVHA), low-valence, low-arousal (HVLA), low-valence, and high-arousal (LVHA) (LVHA). They validated their model against the MANHOB and DEAP datasets and found 81% accuracy. Additionally, several researchers have demonstrated the efficacy of combining ECG and EDA [17]. For valence and arousal state classification, they advocated the use of CNN and LSTM-based models accuracy values for validity (75%) and arousal (76%) were reached. A second study is currently underway in which wearable sensors are being used to make emotion recognition more practical and adaptable across various domains.

2.3. Sensor Fusion Emotion Recognition

Using data from facial expressions and biophysiological markers, the authors of [18] made a promising attempt to predict continuous emotions (potency and arousal). Twenty-seven participants were shown twenty video clips and asked to annotate their emotional states using ten emotional keywords selected from a list of twenty (sadness, happiness, joy, disgust, neutrality, etc.). They collected data on facial expressions while simultaneously recording 4 signals of bio-physiology using six cameras. For obtaining biophysiological signals, expensive equipment and a tightly controlled environment are required. Numerous studies have attempted to circumvent these limitations by identifying emotions using expression data of facial and biophysiological signals, which enables multimodal approaches and non-contact performance measurements, among other things. It has also been able to produce reasonably repeatable results due to recent advancements in the systems of computer vision, big data analytics, and the techniques of deep learning. However, there is still considerable room for improvement in arousal levels. As a result of this development, several researchers have considered combining facial expressions with biophysiological signals [19]. The authors of [20] used electroencephalography and FE as input features to predict potency and arousal. According to the results, their validity accuracy was 75%, and their arousal accuracy was 74%. The authors of [21] predicted potency, arousal, liking, and the four emotional states, as well as the four emotional states themselves, using EEG and facial expressions. They were 54% accurate in their forecasts. Varkalvo and colleagues [22] proposed self-designed methods for estimating the user's emotional state in real-time. They used the EDA, blood volume pressure (BVP), and electroencephalography (EEG) to examine the statistical correlation b/w the emotions experienced and the attributes set of features. Deep convolutional integration is used to classify emotions based on facial expressions, and the classification results are obtained using the FER-2013 dataset: It was determined that they may be classified into seven distinct emotional states by using eight standard classifiers set, with an accuracy of close to 80%. The authors of [23] also used the EDA, facial expressions, electrocardiograms for predicting potency, arousal, liking, and seven different emotional states as input features. To train the model, we used the AMIGOS and datasets of Medical Therapy. The model was then used to predict patients'

emotions receiving anxiety treatment. They achieved a 64% accuracy rate. It was suggested that they combine facial expression data and biophysiological signals in FER using a temporal information retention framework and the AFFDEX SDK, which significantly improved emotion recognition performance. Numerous proposals for multimodal emotion recognition have been made because the use of biophysiological signals and facial expressions enables multimodal approaches [24–28]. Using heart rate variability (HRV) variables, the respiratory rate, and the facial expression. The authors of [29] presented multimodal automatic emotion identification. Each of the 24 participants was shown an image from the international affective picture system (IAPS) for a total of 20 s in order to gauge their emotional response. Participants verbally reported their feelings about the photographs using a self-assessment model and a V-A rating scale. Additionally, they used PSD to analyze the RR interval and SVM and the Cohn-Kanade dataset to categorize facial expressions in the dataset. Using multimodal ER, k predicted validity and arousal by 38.0% and 54.5%, respectively. The authors of [30] also proposed using multimodal ER to predict validity and arousal using FE and EEG signals from a single modality, respectively. This study used external channels with nine facial expressions probability distribution as external channels. For supplementing and completing facial expressions, EEG was employed as an internal channel. These two channels were combined at the feature and decision levels to achieve multimodal ER. The authors of [31] used a migration learning approach to construct a multitask CNN architecture to account for the lack of EEG-facial expression fusion. In EEG detection, two distinct learning targets are identified separately and then combined using diverse support vector machine (SVM) classifiers. Two fusion methods of decision-level based on enumeration weight rules or adaptive enhancement techniques combined facial expressions and electroencephalography. After fusing the DEAP and MAHNOB-Human–Computer interface (abbrv. MAHNOB-HCI) datasets, they achieved 69.74 and 70% validity and arousal accuracy, respectively, and a 6% increase in arousal accuracy. The authors of [32] focused on arousal's recognition and potency. After extracting power spectrum features from EEG signals, they used facial datums in each frame, as features for sequentially detecting valence levels. Compared to biophysiological signals, the use of facial expressions is more versatile and applicable to a broader range of situations. Due to the flexibility and practicability of this approach, active research is being conducted in required real-time emotion recognition areas in a variety of situations.

2.4. Existing Datasets

AffectNet [33] is a dataset containing over a million images of faces gathered from the Internet and 1250 keywords related to emotions in almost six languages. Seven discrete data points of facial expression and information about potency and arousal intensity were annotated onto the collected images. According to the dataset's authors, CK+ [34] contains 593 sequences with seven emotion labels. The EMOTIC [35] classification system includes 26 discrete categories, including continuous dimensional, dominance, and images. EMOTIC includes 23,571 images, some of which were sourced from Google's image search engine. While subjects watch film clips, the ASCERTAIN dataset [36] records their facial images and biophysiological signals. Within 30 s of viewing each clip, subjects are asked to self-report their emotional state via a good ranking. The validity (V) and arousal (A) ratings for each image reflected the user's perceptions' validity. Additionally, their biophysiological data were gathered. This dataset can be used to learn about people's emotions. Numerous datasets, such as the MAHNOB-HCI dataset [37], maintain emotional responses. Eye gaze data and physiological signals are available (EDA, ECG, breathing patterns, and skin temperature). Two experiments revealed several self-reported keywords for arousal, potency, dominance, predictability, and emotion. The Dreamer dataset [38] contains an electroencephalogram (EEG), an electrocardiogram (ECG) signal obtained during the elicitation of audiovisual stimuli [39]. Following each stimulus, 23 participants self-rated their affective state based on potency, arousal, and dominance, with the highest scores indicating the most positive state. Through the use of SVM, this dataset enables

the identification of emotions. Eight physiological signals were collected from 30 different participants for the CASE dataset [40]. The authors of [41] used a dataset to classify the emotional states of drivers into four categories (happiness, irritation, focused attention, and confusion) and collected the data while driving approximately 24 km each. The data set is primarily made up of face videos that have been annotated emotionally by external annotators. The participants were shown a variety of videos to elicit four distinct emotional states: amusing, boring, relaxed, and frightening. They collected data from portable UTDrive DB using smartphones, obviating the need for any specialized equipment. The CIAIR dataset contains driving data of real-world from more than 500 drivers who collectively logged over 60 min of real-world driving [42]. The videos elicited four distinct emotional states. They were instructed to annotate their changing emotions as they progressed through the game using a joystick. Certain datasets include both facial images and biophysiological signals that can classify emotions. The use of driving-related facial image data [43] or biophysiological signals [44] as datasets for emotion recognition is gaining popularity. The data for this study were collected while 77 participants drove the UTDrive DB Classic in real-world urban area highway conditions. Audio, video, the distance between you and the vehicle in front of you, and driver behavior were all recorded and incorporated into the data collection process. This document contains no sentiment annotations. The data is comprised of three-channel videos from three cameras, three-channel audio from three cameras, GPS and control signals. Numerous biophysiological signals (ECG, EDA, EMG, and RESP) are collected and analyzed in the Drive dataset under various stress conditions. Signals are collected for 50–90 min at a time while driving.

3. Methodology

According to the literature review, significant research has been conducted on real-time and offline emotion recognition based on facial expressions using benchmark datasets, both in real-time and without the use of a network. We can see the overall structure of our research framework in the given Figure 1. As a result of this slew of other challenges and limitations, the first and most significant contribution task in this research is the development of a driver's emotion recognition (DER) dataset. Individual high-resolution RGB images of each subject's seven basic expressions will be captured, and each subject will contribute to developing a facial expression image dataset containing these images. The second phase will be devoted to developing a face detection system. The proposed detector will be tailored specifically for car indoor face detection to be effective and precise in its operation and efficient in its operation.

Additionally, the developed face detection system will be highly robust and perform well under adverse conditions such as low lighting and facial obstructions. Transfer learning is a cutting-edge technique for developing a custom dataset model from scratch in FER. The purpose of this research is to develop the most advanced and highly accurate convolutional neural network model capable of recognizing driver emotions in a vehicle interior scene. We will use holdout cross-validation, also known as %age splitting, to partition the benchmark and our developed dataset. The dataset will be divided into two parts: a trainset (which will be randomly split 70% of the time) and a test set (which will be randomly split 30% of the time). The accuracy and F-measure metrics used for evaluation in the proposed model's performance in this manuscript in the final section, will be discussed in greater detail later.

3.1. Datasets

The datasets used in this study are defined in elements in this section. Visual data from the FER image dataset is the primary basis for the computer execution of FER. Most researchers working in the FER domain rely on existing facial expression datasets due to funding constraints, time, effort, and algorithm performance evaluation requirements. CK+, JAFFE, FER-2013, RAF-dB, and AffectNet facial expression datasets are the most widely used. The CK+38 and JAFFE datasets, which were chosen for evaluation in this

study, are currently the two most widely used standard datasets from the very beginning of expression research to evaluation. The FER-2013 dataset, the most widely used dataset for FER available for public use, was also used in this study. Figure 2 and Table 1 show the three most widely used DFE datasets used in this work, and the distribution of every class in the datasets is discussed.

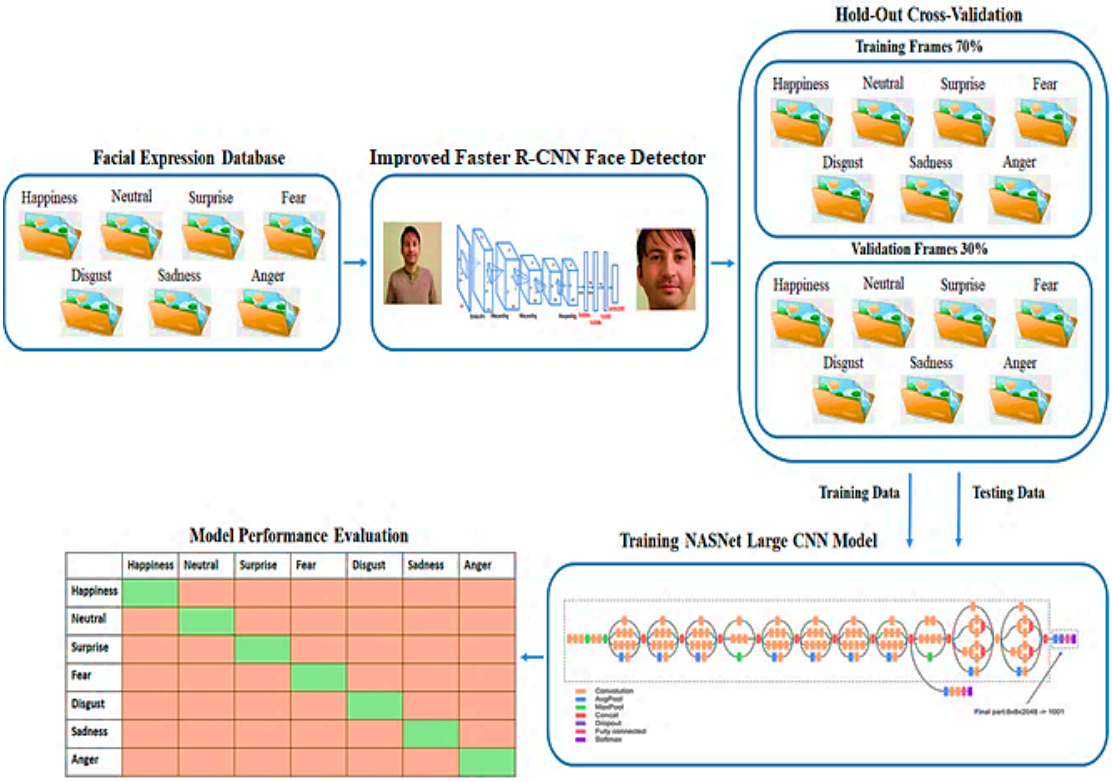

Figure 1. Proposed image-based driver emotion recognition framework.

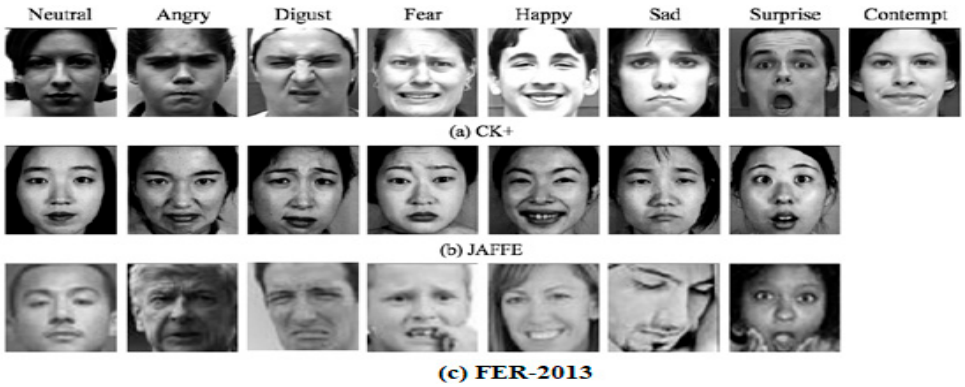

Figure 2. Some random samples from CK+, JAFFE, and FER-2013 facial expressions datasets.

Table 1. The distributions of every class in each dataset.

	FER-2013	CK+ (Last Frame)	JAFFE	AffectNet	Custom
Angry	4593	45	30	25,382	460,164
Disgust	547	59	30	4303	444,819
Fear	5121	25	31	6878	274,475
Happy	8989	69	31	134,415	38,885
Sad	6077	28	31	25,959	36,330
Surprise	4002	83	30	14,590	29,995
Neutral	6198	327	30	75,376	Nil
Contempt	0	18	0	4250	279,246

3.1.1. The Japanese Female Facial Expression (JAFFE) Dataset

Basic facial expressions (FE) for each of the three datasets. With the addition of a contempt class in CK+, each dataset contains seven fundamental expressions which are commonly used such as: happy, angry, disgusted, fear, sad, contempt, and surprised. The FER-2013 dataset consists of wild-type facial expressions, whereas the CK+ and JAFFE datasets contain posed and collected FE in a laboratory setting.

3.1.2. The Extended Cohn-Kanade Dataset (CK+)

In 2010, the CK dataset was expanded [16], resulting in a 22% and 27% increase in sequences and subjects, respectively. The dataset consists of 327 images in sequences of postural and non-postural (spontaneous) expressions ranging from neutral to highest emotions, digitized to 640 × 480 and labeled with FACS-encoded emotion labels for the highest edges. The dataset's 123 subjects age approximately from 18 to 50 years old (81% European-American, 13% African-American, and 6% other races), with females accounting for 69% of the total. The dataset includes a category of "contempt" in adding to the basic seven FE, for a total of eight facial expressions. There are also some baseline evaluation results and methods for tracking contents and presence features and emotion and AU labeling in this dataset.

3.1.3. FER-2013 Dataset

The ICML 2013 Representation Learning Challenge [14] was the first to introduce the database for the recognition of facial expressions 2013 (FER2013). The collection contains 35,887 pictures with a resolution of 48 by 48 pixels, the majority of which were taken in the field. There were 28,709 photographs in the training set, 3589 images in the test set, and 3589 in the test set. Initially, faces in the Google Image Search API database were automatically captured. One of six fundamental expressions or a neutral expression was then applied to the faces. The FER is more common in partial faces, low-contrast pics, shows, and facial occlusion than in the other datasets.

3.1.4. AffectNet Dataset

AffectNet is a new dataset of real-life FE created by gathering and annotating facial images. AffectNet is an FE dataset of over 1 million facial images collected from the Internet by querying with the major three search engines with 1250 emotions from six different languages people related keywords. The presence of different seven FE (categorial model) and the amount of valence and stimulations were manually explained in about half-of-the retrieved images (440,000) (dimensional model). AffectNet is the most publicly available dataset of FEs, valence, and stimulation, allowing researchers to investigate automated FER in two different emotion models. In the categorical model, the two baselines are used to predict the intensity of valance and simulation to classify images by deep neural networks. Our deep neural network baselines outperform the conventional machine learning approach and off-the-shelf FER method on various evaluation metrics.

3.1.5. Custom Dataset

A benchmark dataset is created as part of a static camera was mounted next to the car's front screen to record videos of drivers making various facial expressions to capture their emotions on film. The photographs were taken from a range of vehicles, including a Toyota Prius, a Honda Civic, and a Toyota Landcruiser. Each subject was recorded for ten minutes in each vehicle, followed by manual separation and labeling of the recordings. A high-resolution camera is used to create the benchmark dataset for emotion recognition. Thirty subjects will be recruited to participate in the dataset development phase to collect image data. Each subject (driver) in the study is a male between the ages of 25 and 40, wearing glasses or not wearing glasses, wearing a cap or not wearing a cap, and with or without a beard. The videos are recorded and analyzed in a moving car interior scene, where obstacles and significant lighting changes interfere with the face detection and

emotional recognition systems. To demonstrate the benchmark driver emotions dataset's effectiveness, the proposed deep learning model is trained on some benchmarks and a custom-developed dataset.

3.2. Face Detection System

Object detection is the face as a process of classifying and identifying the contents in an image. The in-depth learning method, R-CNN, combines the rectangular region schemes with CNN features. R-CNN is a two-step detection procedure. The first step recognizes a subset of regions in an image that might contain an object. The second step organizes the object in each region.

Our proposed face detector is called improved faster R-CNN, which is composed of two schemes. The first scheme is a deep, fully convolutional network that recommends regions, and the second module is the improved fast R-CNN detector that uses the proposed areas. Figure 3 shows examples of face detection using improved faster R-CNN, the object detection in an image in the system is a single, unified network. The PRN approach expresses the improved faster R-CNN module, using the most recently advanced language for the neural networks with attention appliances.

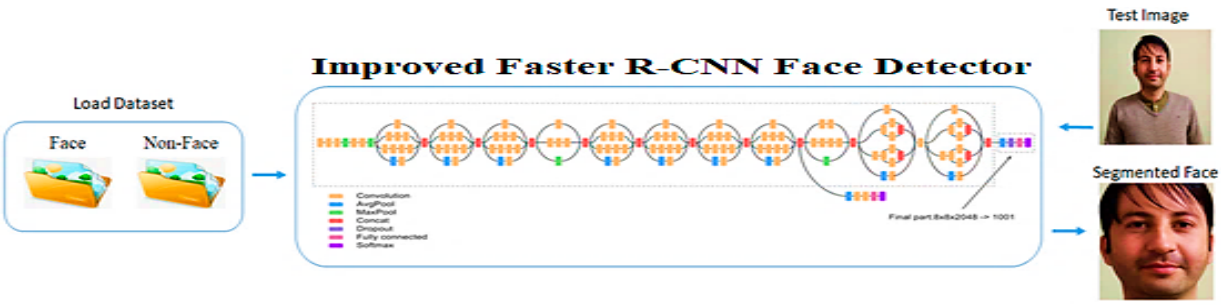

Figure 3. Face detection using improved faster R-CNN model.

3.3. Regional Convolutional Neural Network (R-CNN)

The R-CNN detector first generates an area in the image schemes using an algorithm such as Edge Boxes. The proposed scheme crops the images and resizes them. Then, the CNN organizes the gathered and resized regions in the image. The CNN features were trained by a support vector machine (SVM) that the area of the image in the proposed scheme is bounded to be fixed are identified shown in Figure 4.

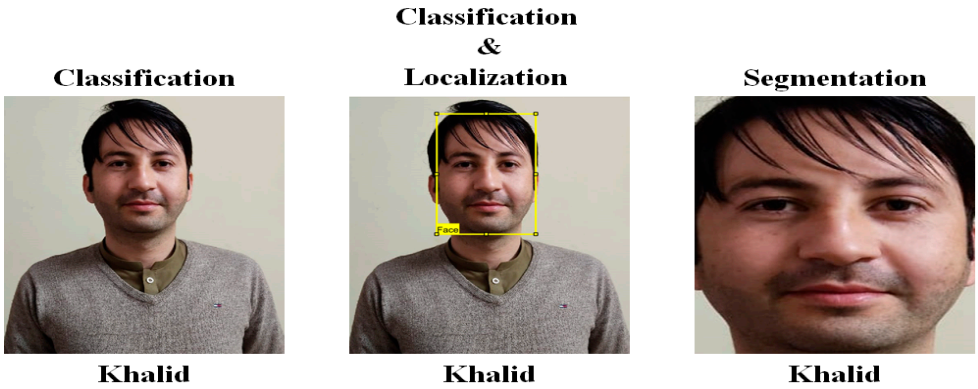

Figure 4. Difference between classification, localization, and segmentation.

3.4. Data Augmentation

Data augmentation was used in different perspectives in computer vision. Data augmentation increases the images (Figure 5) in the dataset by replicating it using multiple image processing techniques. There are many common data augmentation techniques were

used for images such as scaling, cropping, flipping, sharpening, blueness, noise removal, noise addition, contrast adjustment, rotation, translation, affine transformation, RGB color shifting, etc.

Figure 5. A single image replicated using different image processing techniques.

3.5. Transfer Learning Based Driver Emotion Recognition (DER)

Transfer learning is an in-depth model that is trained for one-to-many tasks in a learning approach. A fine-tuning transfer learning is faster and easier than training in a network from scratch. The transfer learning approach, as distributed over a considerable extent technique, enables the researcher to train the models using related small, labeled data by leveraging mostly used models that have been trained on a large dataset. The transfer learning approach is mostly used in computer vision for object detection in images, image recognition, speech recognition, and other applications. For full training a cycle on the whole dataset, a new model would be required because in transfer learning the model did not need to be trained for several epochs. It can histrionically decrease training time and compute possessions.

3.6. Transfer Learning

The transfer learning approach commonly uses the following process steps, the workflow of the transfer learning workflow, as shown in Figure 6.

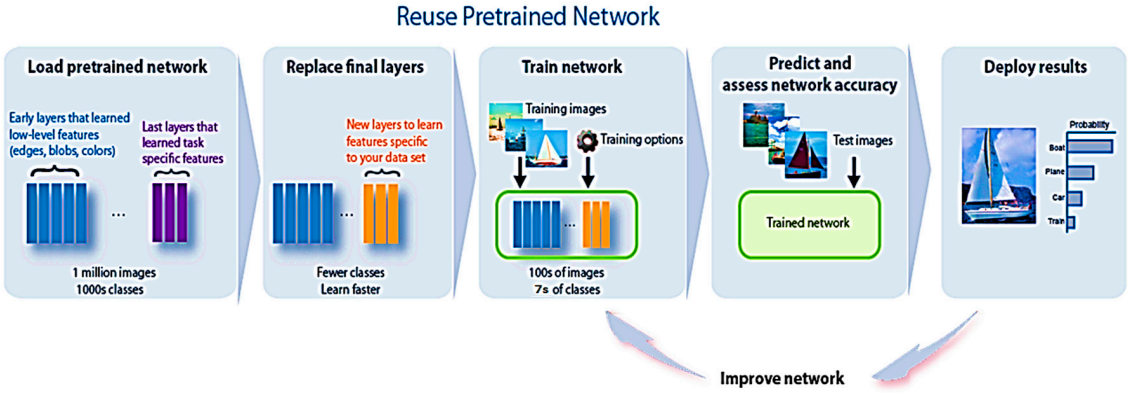

Figure 6. Transfer learning workflow.

Load a pre-trained network for the new task and select the most relevant pre-trained network for the most similar task.

The classification layers will be changed for the new task, because you may also select a fine tune, the weight depends upon the data availability of the new task. Moreover, for many data, you have to use more layers to select a fine tune, but for a smaller number of data, the fine tuning may lead to an over-fitted model.

For a new task, the network must be trained on the data.

After training the network, check the accuracy of the newly trained network.

3.7. Transfer Learning in Pre-Train Convolutional Neural Networks

3.7.1. NasNet Large

To find the best convolutional architectures for a given dataset, we use search algorithms. Neural architecture search (NAS) is the major search method that we want to deploy. Child networks with varied architectures are sampled by a controller RNN in NAS. Child networks are taught to achieve some accuracy on a validation set that is held out for convergence. The resulting accuracy values are used for updating the controller, which in turn generates more accurate architectures over time. The policy gradient is used for updating the controller weights.

A new search space is designed that allows the best architecture discovered on the CIFAR-10 dataset to be scaled up to larger, higher-resolution image datasets in a range of computing environments. Based on a realization that architecture engineering with CNNs usually uncovers recurring patterns that include combinations of convolutional filter banks and nonlinearities as well as a careful selection of connections, the NasNet search space was created (for example, the repeated modules in the inception and ResNet models). These findings suggest that the controller RNN may be able to predict a generic convolutional cell that is displayed in these patterns. To accommodate inputs of any spatial dimension and depth of filtering, this cell can be stacked in a sequence. In our method, the convolutional nets' overall designs are predetermined manually. They are made up of convolutional cells that have the same shape as the originals but are weighted differently. Two types of convolutional cells are used to quickly develop scalable architectures for images of any size: (1) convolutional cells that produce a feature map with that of the same dimension, and (2) convolutional cells that return a feature map with a two-fold reduction in height and width.

The first and second types of convolutional cells are referred to as Normal Cell and Reduction Cell, respectively. The initial operation applied to the cell's inputs is given a two-step stride to minimize the cell's height and width. Our convolutional cells support striding since they take all operations into account when constructing them. The Normal and Reduction Cell structures that the controller RNN seeks are different in convolutional nets. The following search area can be used to look for cell shapes. There are two hidden states, h I and h (i-1), available for each cell in our search space. To begin, these initial hidden states are the results of two cells in the preceding two lower layers or the input image, respectively. The controller RNN makes recursive predictions about the rest of the convolutional cell structure based on these two initial hidden states (Figure 7). The controller predictions for each cell are organized into B blocks, with each block consisting of five prediction steps performed by five distinct SoftMax classifiers representing discrete choices of block elements.

Step 1. Select a hidden state from h i, h (i-1), or the set of previously created hidden states.

Step 2. From the same options as in Step 1, select a second hidden state.

Step 3. In Step 1, choose the hidden state you want to apply an operation upon.

Step 4. After selecting a hidden state in Step 2, select an operation to apply to it.

Step 5. Determine how the outputs from Steps 3 and 4 will be combined for creating a new hidden state.

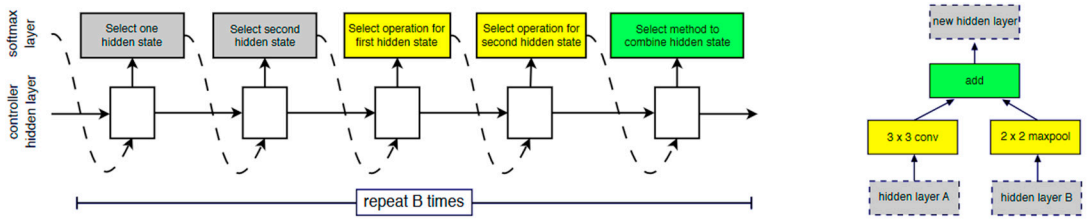

Figure 7. A recursive block of neural architectural search network used for features learning.

It is possible to use the newly created hidden state as an input in the following blocks. For each of the B blocks in the convolutional cell shown in Figure 7, the controller RNN propagates over the five previous predictions B times. However, due to computing limitations, we were unable to fully explore this region in our studies and concluded that B = 5 provides satisfactory results.

The controller RNN selects an operation to apply to the hidden states in steps 3 and 4. We compiled the following list of operations (Table 2) based on their frequency of occurrence in the CNN literature:

Table 2. NasNet layer architecture.

Identity	1×3 Then 1×3 Convolution
1×7 then 1×7 convolution	3×3 dilated convolution
3×3 average pooling	3×3 max pooling
5×5 max pooling	7×7 max pooling
1×1 convolution	3×3 convolution
3×3 depthwise-separable conv	5×5 depthwise-separable conv
7×7 depthwise-separable conv	Last fully connected

3.7.2. Features Weights Optimization

Detailed information about how to train using our clinical dataset to fine-tune pre-trained CNNs, we examined each one separately. As a result, we used NasNet's large pre-trained CNNs. Our study makes use of image augmentation. Additionally, to the initial dataset, another dataset of training is generated using the data augmentation technique. During training, data augmentation can be used to mitigate the overfitting problem associated with deep CNNs. We applied a random horizontal and vertical shift to an extent of 10% of the original dimensions in this study. Additionally, random rotation (20°) was applied to the training images, along with a small random zoom. Additionally, we flip the images horizontally to increase the dataset's size. To fine-tune all networks, we removed all fully connected layers and used only the convolutional portion of each model's architecture. We added a global average pooling layer on top of the final convolutional layer, followed by a final classification layer that makes use of SoftMax non-linearity. With a learning rate of 0.0001 and a momentum of 0.9, we employed stochastic gradient descent (SGD) optimization for 50 iterations to fine-tune the networks. The loss function was categorical cross-entropy in all situations. It is used to adjust the hyperparameters in the validation set. To be clear, each network's input is unique in terms of its size. The initial stage in data preparation was to resize and store all photos in various files based on the varied sizes of model inputs that were used. Table 2 shows the comparison of pre-trained CNN models. Both models were trained with identical initialization and learning rate rules.

4. Experiments

Facial expression recognition methods such as the one proposed here have been tested and proven to be effective on various standard datasets, which were used in the development of this section. A thorough comparison of the proposed technique with current FER techniques and quantitative and qualitative evaluations of the results performed on the data collected are also included in the study. A more specific example is that the proposed

system uses two reference datasets. Each dataset in the proposed system is divided into training and testing sets at random, with the training set being significantly larger than the testing set and the training set being also larger than the testing set. As previously stated, to conduct experiments, it is necessary to change the number of training and testing images. All simulations are performed using the MATLAB R2021a simulation platform included in the proposal. All of this is accomplished on a workstation PC equipped with dual Xeon processors, 48 GB DDR4 RAM, which are running the Windows 10 operating system. The following sections provide a more detailed (Table 3) explanation of each of the two datasets (RAF and AffectNet) used in the experiment.

Table 3. Specification of GPU using for model training.

Manufacturer	Nvidia
Model	RTX 2080Ti
Memory	11 GB GDDR-6
Cores	4352
TMUS	272
ROPS	88
Bus Width	352 bit

4.1. Experiments on the JAFFE Dataset

A randomized hold-out splitting procedure is used to conduct experiments on the JAFFE dataset, which allows for the most accurate results. In the first phase, 189 images (70% of the total) are used as a training set, and 24 images (30% of the total) are used as testing images. In the second phase, 189 images (70% of the total) are used as training images. After that, an additional 6237 training images are added to the JAFFE augmented dataset then used to validate the model. A total of 792 validation images are also used in the process of model validation. The accuracy achieved using the original JAFFE dataset is 88.82% in Figure 8a while Figure 8b shows the NasNet-Large CNN model training and validation loss plot for the JAFFE dataset. The given Figure 8a,b illustrates the contrast accuracy and loss for the proposed NasNet-Large model in terms of direct proportion with the epochs.

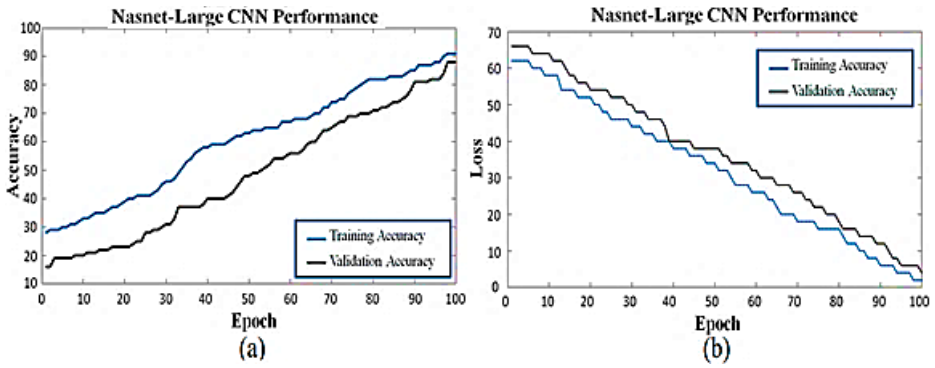

Figure 8. (**a**) NasNet-Large CNN model training and validation accuracy plot for JAFFE original dataset, (**b**) NasNet-Large CNN model training and validation lossplot for JAFFE dataset.

The accuracy is achieved using the augmented JAFFE dataset which is 98.48% in Figure 9a whereas Figure 9b shows the NasNet-Large CNN model training and validation loss plot for the JAFFE augmented dataset. The given Figure 9a,b illustrates the contrast accuracy and loss for the proposed Nasnet-Large model for the augmented dataset in terms of direct proportion with the epochs.

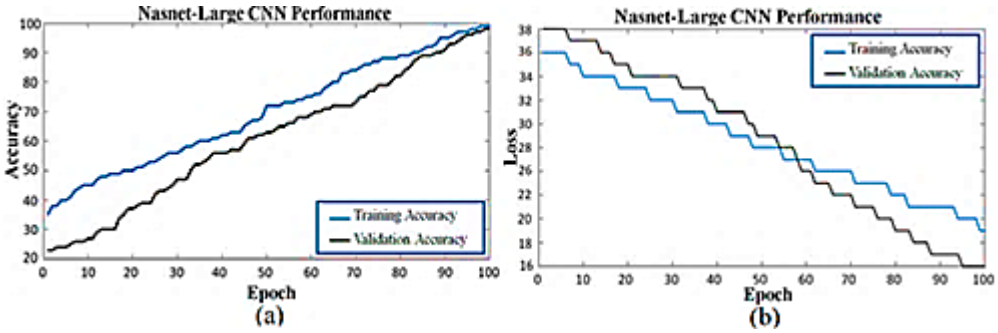

Figure 9. (**a**) NasNet-Large CNN model training and validation accuracy plot for JAFFE augmented dataset, (**b**) NasNet-Large CNN model training and validation lossplot for JAFFE augmented dataset.

4.2. Experiments on the CK+ Dataset

Experiments on the CK+ dataset are conducted using randomized hold-out splitting. The first phase uses 444 images (70% of the total) as a training set and 192 images (30% of the total) as testing images. The second phase adds 14,652 training images and 6336 validation images to the CK+ augmented dataset, bringing the total number of training images to 14,652. The accuracy obtained using the original CK+ dataset is 95.97% in Figure 10a,b shows the model training and validation loss plot for CK+ Dataset.

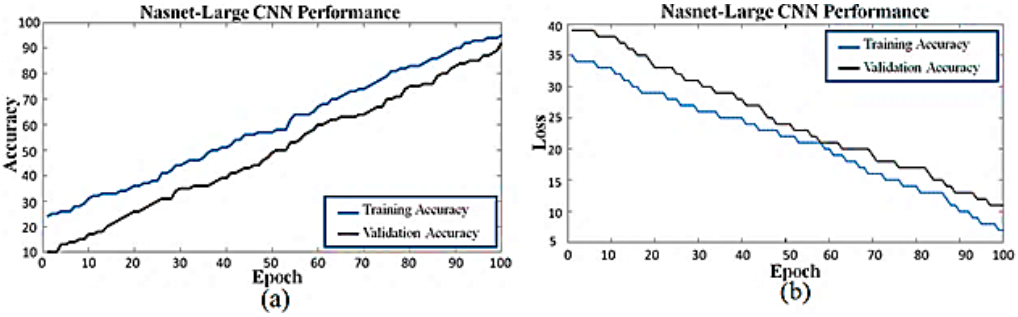

Figure 10. (**a**) NasNet-Large CNN model training and validation accuracy plot for CK+ original Dataset, (**b**) NasNet-Large CNN model training and validation lossplot for CK+ Dataset.

The accuracy is obtained using the augmented CK+ dataset which is 99.73% in Figure 11a,b shows the model training and validation loss plot for the CK+ augmented dataset.

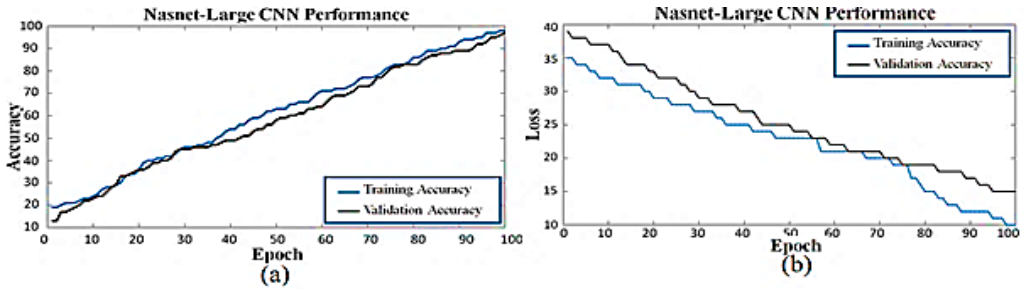

Figure 11. (**a**) NasNet-Large CNN model training and validation accuracy plot for CK+ Augmented Dataset, (**b**) NasNet-Large CNN model training and validation lossplot for CK+ Augmented Dataset.

4.3. Experiments on the FER-2013 Dataset

Experiments on the FER-2013 dataset are performed using randomized hold-out splitting. In the first phase, 23,569 (70%) images are used as a training set and 10,658 (30%) images are used for testing purposes. In the second phase, the FER-2013 dataset is augmented where the number of training images is 777,777 and 321,691 images that are used

for model validation. The accuracy of the original FER-2013 dataset is achieved by the NasNet-Large model which is 98.34% shown in Figure 12a,b for the model training and validation loss plot.

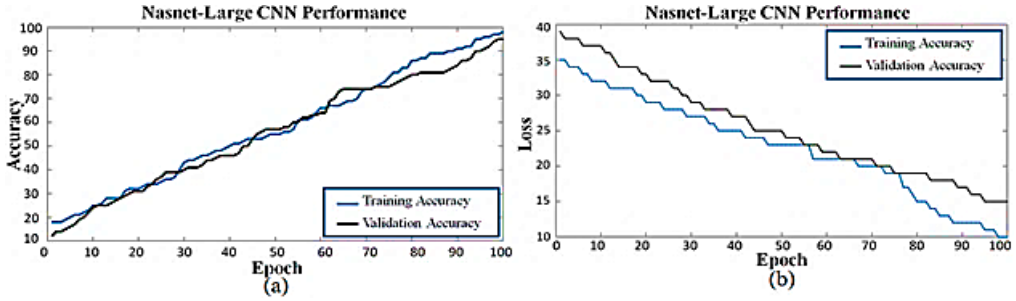

Figure 12. (**a**) NasNet-Large CNN model training and validation accuracy plot for FER-2013 dataset, (**b**) NasNet-Large CNN model training and validation lossplot for FER-2013 dataset.

The accuracy of the augmented FER-2013 dataset is achieved by the NasNet-Large model which is 99.95% in Figure 13a while Figure 13b shows the model training and validation loss plot.

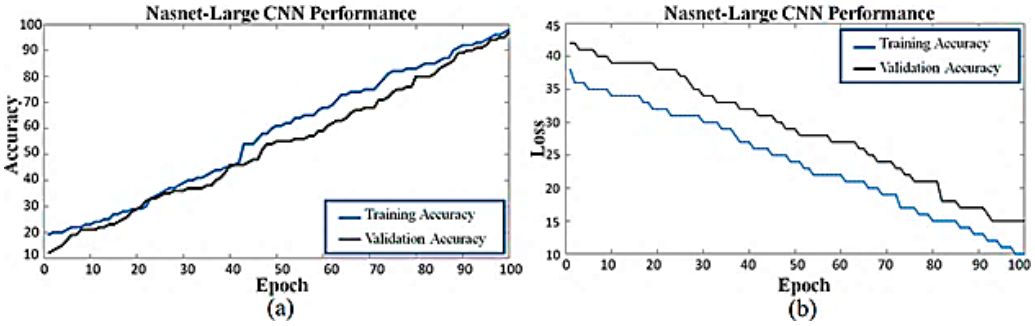

Figure 13. (**a**) NasNet-Large CNN model training and validation accuracy plot for FER-2013 augmented dataset, (**b**) NasNet-Large CNN model training and validation lossplot for FER-2013 augmented dataset.

4.4. Experiments on the AffectNet Dataset

Experiments on the AffectNet dataset are performed using randomized hold-out splitting. In the first phase, 187,807 (70%) images are used as a training set and 87,346 (30%) images are used for testing purposes. In the second phase, the AffectNet dataset is augmented where the number of training images is 2,817,105 and 1,310,190 images that are used for model validation. The accuracy is achieved by the original AffectNet dataset is 80.99% in Figure 14a whereas Figure 14b shows the training and validation loss plot.

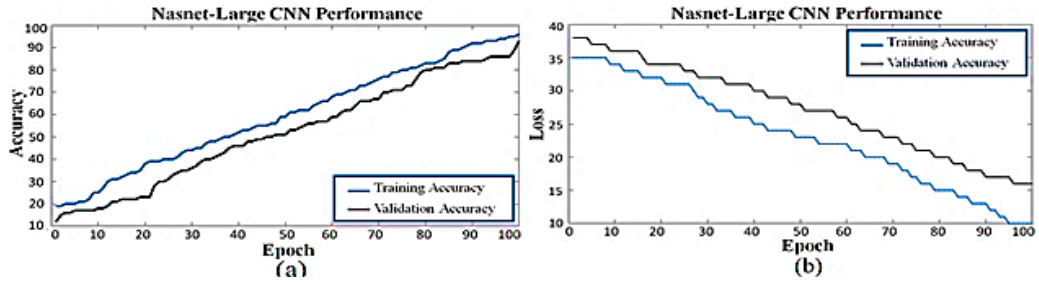

Figure 14. (**a**) NasNet-Large CNN model training and validation accuracy plot for AffectNet dataset, (**b**) NasNet-Large CNN model training and validation lossplot for AffectNet dataset.

The accuracy is achieved by the augmented AffectNet dataset which is 95.28% as shown in Figure 15a,b—the training and validation loss plot.

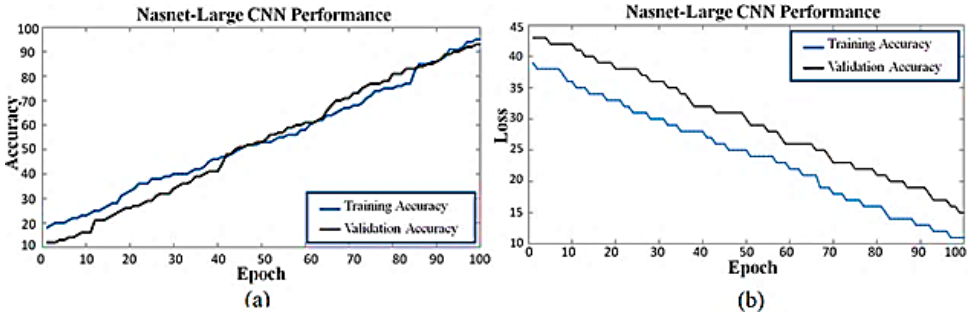

Figure 15. (**a**) NasNet-Large CNN model training and validation accuracy plot for AffectNet augmented dataset, (**b**) NasNet-Large CNN model training and validation lossplot for AffectNet augmented dataset.

4.5. Experiments on the Custom Dataset

Experiments on the custom dataset are performed using randomized hold-out splitting. In the first phase, 763,880 (70%) images are used as a training set and 329,926 (30%) images are used for testing purposes. In the second phase, the custom dataset is augmented where the number of training images is 5,347,160 and 2,309,482 images that are used for model validation. The accuracy of the original custom dataset is achieved by the NasNet-Large CNN model which is 97.65% as shown in Figure 16a,b—the training and validation loss plot.

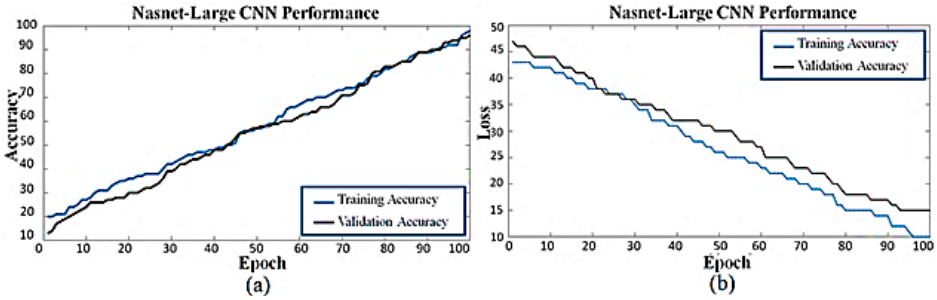

Figure 16. (**a**) NasNet-Large CNN model training and validation accuracy plot for Cusotm dataset, (**b**) NasNet-Large CNN model training and validation lossplot for Custom dataset.

The accuracy of the augmented custom dataset is achieved by the Nasnet-Large CNN model which is 99.15% as shown in Figure 17a,b—the training and validation loss plot.

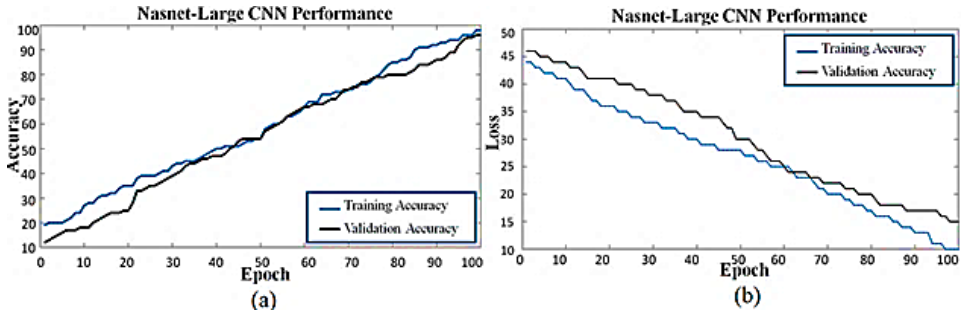

Figure 17. (**a**) NasNet-Large CNN model training and validation accuracy plot for Custom augmented dataset, (**b**) NasNet-Large CNN model training and validation lossplot for Custom augmented dataset.

5. Performance Comparison with State-of-the-Art

We show the performance of the model proposed on the above data sets. The model is also accomplished, verified in a validation set, and the accuracy of the test set presented in each case, in part of the data set. Before looking at the performance of the model in multiple data sets, we discuss our training approach briefly. For every collection of data, we trained a model in our experiments but tried to preserve its architecture and hyperparameters. We employed the Adam optimizer at a base learning rate of 0.005. (a different optimizer was used, including stochastic gradient descent, and Adam appeared to be slightly more successful). With the weight decay of 0.001 value, L2 regularization was also added. On the FER datasets, training our model took about 2 h while the JAFFE dataset consists of fewer images, so the model training time was less than 10 min. We used oversampling for classes with fewer images in the dataset for model training, which also resolved the imbalanced data problem, leading to model generalization. Classes could have the same order for training the model on colossal no. For images and training models against small invariant transformation, data augmentation is used in training sets. To improve the data, we use flips, small rotations, and minor distortions.

Other data sets for the recognition of FE are more accessible in use than the FER-2013 dataset. Aside from the variation of intra-class in FER, the unbalanced nature of the different emotion classes is another major challenge in this dataset. Happiness and neutrality have many more examples than other expressions. The model was trained using 28,709 images from the training sets, verified with 3500 validation images, and tested with 3589 images from the test set to determine model quality. On the test set, we were able to achieve 79.1% accuracy on the actual size dataset and for the augmented dataset the model accuracy is 99.95%. The year 2013 is determined by comparing our model results with some previous work on FER Table 4 and the accuracy graph (Figure 18).

Table 4. FER 2013 dataset's classification accuracy.

Methods	Accuracy Rate
Unsupervised classification [38]	66.1%
Bag of visual words [39]	63.7%
VGG features +SVM classifier [24]	67.43%
Transfer learning in GoogleNet [40]	64.1%
Facial expression recognition on SoC [41]	65.9%
Mollahosseini et al. [10]	65.9%
Transfer learning in VGG (Aff-Wild) [42]	76.01%
Proposed model (Original dataset)	98.34%
Proposed model (augmented dataset)	99.95%

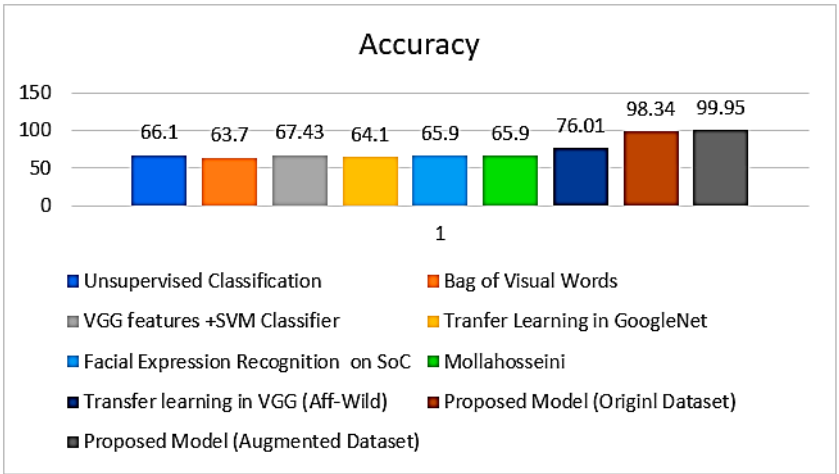

Figure 18. FER 2013 dataset's classification accuracy.

For training, we have used the total amount of the dataset of one hundred and twenty (120) images. From this viewpoint, we have used twenty-three (23) images for validation, and seventy images (70) for JAFFE dataset testing. The total accuracy of this dataset is 92.8%. The proposed model results have been achieved and have been compared with the results from some state-of-the-art work which are given in Table 5 and the accuracy graph (Figure 19).

Table 5. JAFFE dataset's classification accuracy.

Method	Accuracy Rate
Mixing image components LBP + ORB [43]	87.82%
Fisherface [44]	90.1%
CNN + HOG features [45]	89.71%
Saliency face map patch [46]	92.2%
Learnable features + multi-class SVM [47]	96.41%
Proposed model (original dataset)	88.82%
Proposed model (augmented dataset)	98.48%

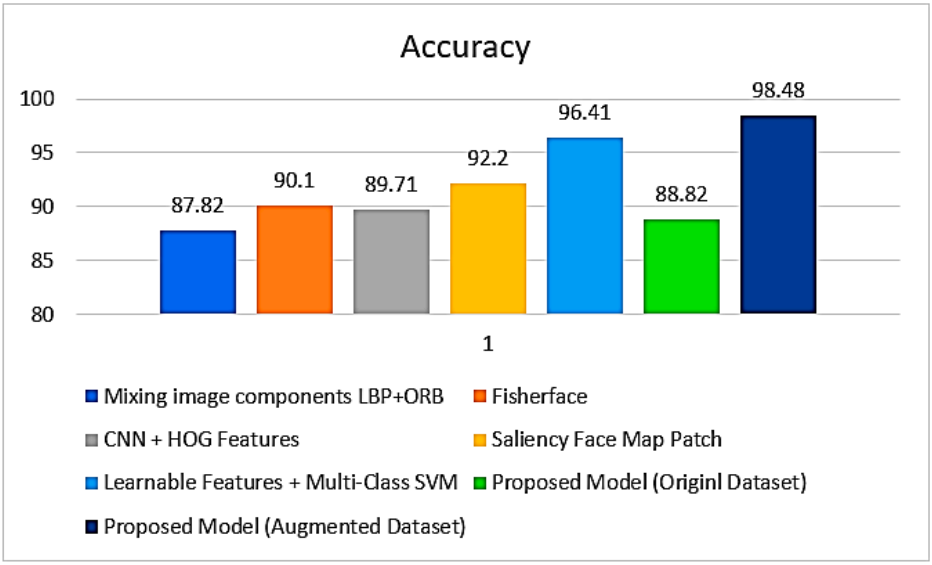

Figure 19. JAFFE dataset's classification accuracy.

This illustration signifies the proposed dataset which also consists of the same dataset with the range of 120 in the total amount of images to be trained and tested under the simulation environment in comparison with the state-of-the-art solution.

From this viewpoint, we have used twenty-three (23) images for validation, and seventy images (70) for CK+ dataset testing. The total accuracy of this dataset is 92.8%. The proposed model results are achieved and have been compared with the results from some state-of-the-art work which are given in Table 6 and the accuracy graph (Figure 20).

Table 6. CK+ dataset's classification accuracy.

Method	Accuracy Rate
DNN [29]	68.26%
Aam-SVM [44]	68.26%
DRL-CNN [45]	82.86%
PGC [46]	62%
FPD-NN [47]	66.36%
Proposed model (original dataset)	95.97%
Proposed model (augmented dataset)	99.73%

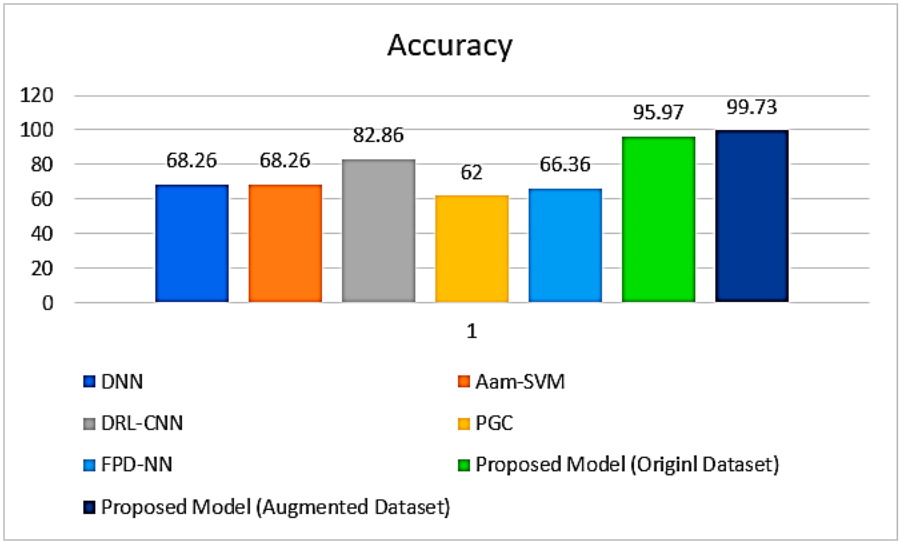

Figure 20. CK+ dataset's classification accuracy.

For validation and AffectNet training and testing, the same dataset was used in the range of 120 images to be contrasted with the existing works.

From this scenario, we have used twenty-three (23) images for validation, and seventy images (70) for AffectNet dataset testing. The total accuracy of this dataset is 92.8%. The proposed model results have been achieved and are compared with the outcomes from some state-of-the-art work which are given in Table 7 and the accuracy graph (Figure 21).

Table 7. AffectNet dataset's classification accuracy.

Method	Accuracy Rate
RAN [31]	59.50%
VGG-16 [4]	51.11%
GAN-Inpainting [34]	52.97%
DLP-CNN [16]	54.47%
PG-CNN [18]	55.33%
ResNet-PL [26]	56.42%
OAD-CNN [final]	61.89%
Proposed model (original dataset)	80.99%
Proposed model (augmented dataset)	95.28%

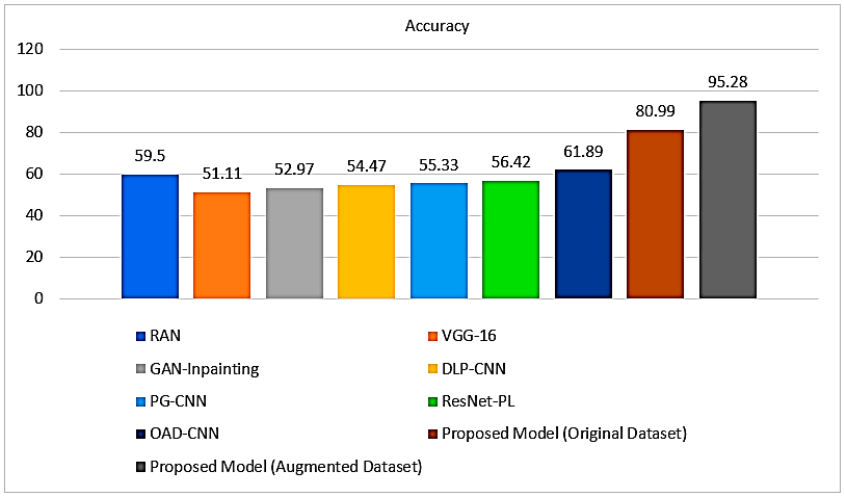

Figure 21. AffectNet dataset's classification accuracy.

Applications of DFERS

A new approach in which the facial expressions of the drivers can be identified and witnessed through advanced artificial deep learning models was established. This approach has the ability to detect the expression of the driver while driving in which there are multiple expressions such as anger; sad, happy, surprised, disgusted, natural, etc. types of moods can be detected by using advanced image processing modules to train and test the object. It works in such a way in which there will be multiple cameras installed inside the vehicle which will be monitoring the condition of the driver and from that viewpoint, these AI-enabled cameras will have the ability to capture the moment of the real-time scenario and will report immediately to the base station or to other emergency stations connected with the vehicles' network. The major applications of the DRER systems are increasing with rapid speed and demand is also increasing as sometimes the drivers become lazy and this AI scheme gives them a signal when it detects some suspicious activity from the driver, such as the driver is about to nod off or something terrible is about to happen to him regarding fainting. Then, this scheme generates an alarm and gives a signal to the driver as well as to other connected base station focal individuals.

6. Conclusions

This paper proposed an algorithm for recognizing the emotional state of a driver. The deep learning algorithm model uses driver face images to identify the driver's emotional (DE) state. According to the proposed FER model, it is possible to identify a DFE state without requiring the driver to perform any additional efforts. The faster RCNN model is improved which is used for the detection of the driver face region, the features learning block of the faster RCNN is replaced with a custom CNN block which improves the face detection accuracy and efficiency. Transfer learning in the NasNet Large CNN model is performed by replacing the ImageNet data with the custom-created dataset of driver emotions; the CNN model used for the recognition of facial expression consists of 1243 layers. The custom-created dataset has seven basic driver emotions. The proposed face detection and facial expression recognition models have been evaluated using a custom dataset; the effectiveness of the proposed model can be analyzed with achieving the high accuracy. The proposed model is also evaluated using some benchmark facial expression recognition datasets, i.e., JAFEE, CK+, FER-2014, and AffectNet. Using the benchmark datasets, the proposed model outperformed some state-of-the-art facial expression recognition models. The proposed model is efficient and accurate and can be deployed by various hardware to recognize the driver's emotions and enhance the performance of an automatic driver assistance system.

Author Contributions: K.Z. and S.Z. conceptualized the study and obtained the resources; K.Z. and S.Z. performed data curation and administrated the project; K.Z. and S.Z. performed formal analysis; K.Z. and S.Z. prepared methodology, investigated the study, and wrote the original draft preparation; S.Z. was responsible for funding acquisition; K.Z. and S.Z. validated the study; K.Z., S.Z., S.M.S., M.S., P.L. and A.H. reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National key research and development program, grant number 2018YFB1600202 & 2021YFB1600205; National Natural Science Foundation of China, grant number 52178407.

Data Availability Statement: No external data were used.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Albentosa, J.; Stephens, A.N.; Sullman, M.J. Driver anger in France: The relationships between sex, gender roles, trait and state driving anger and appraisals made while driving. *Transp. Res. Part F Traffic Psychol. Behav.* **2018**, *52*, 127–137. [CrossRef]
2. FakhrHosseini, S.; Ko, S.; Alvarez, I.; Jeon, M. Driver Emotions in Automated Vehicles. In *User Experience Design in the Era of Automated Driving*; Springer: Cham, Switzerland, 2022; pp. 85–97.

3. Nakisa, B.; Rastgoo, M.N.; Rakotonirainy, A.; Maire, F.; Chandran, V. Automatic Emotion Recognition Using Temporal Multimodal Deep Learning. *IEEE Access* **2020**, *8*, 225463–225474. [CrossRef]
4. Lu, C.; Zheng, W.; Li, C.; Tang, C.; Liu, S.; Yan, S.; Zong, Y. Multiple spatio-temporal feature learning for video-based emotion recognition in the wild. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018; pp. 646–652.
5. Chung, W.-Y.; Chong, T.-W.; Lee, B.-G. Methods to Detect and Reduce Driver Stress: A Review. *Int. J. Automot. Technol.* **2019**, *20*, 1051–1063. [CrossRef]
6. Chang, W.Y.; Hsu, S.H.; Chien, J.H. FATAUVA-Net: An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 17–25.
7. Kollias, D.; Zafeiriou, S. A multi-task learning & generation framework: Valence-arousal, action units & primary expressions. *arXiv* **2018**, arXiv:1811.07771.
8. Theagarajan, R.; Bhanu, B.; Cruz, A. Deepdriver: Automated system for measuring valence and arousal in car driver videos. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 2546–2551.
9. Pavlich, C.A. A Cold Encounter: The Effects of Aversive Stimulation on Verbal and Nonverbal Leakage Cues to Deception. Ph.D. Thesis, The University of Arizona, Tucson, AZ, USA, 2018.
10. Stel, M.; van Dijk, E. When do we see that others misrepresent how they feel? detecting deception from emotional faces with direct and indirect measures. *Soc. Influ.* **2018**, *13*, 137–149. [CrossRef]
11. Bruni, V.; Vitulano, D. SSIM based Signature of Facial Micro-Expressions. In Proceedings of the International Conference on Image Analysis and Recognition, Póvoa de Varzim, Portugal, 24–26 June 2020; Springer: Cham, Switzerland, 2020; pp. 267–279.
12. Oh, Y.H.; See, J.; Le Ngo, A.C.; Phan, R.C.W.; Baskaran, V.M. A survey of automatic facial micro-expression analysis: Datasets, methods, and challenges. *Front. Psychol.* **2018**, *9*, 1128. [CrossRef]
13. Prasanthi, T.L. Machine Learning-based Signal Processing by Physiological Signals Detection of Stress. *Turk. J. Comput. Math. Educ.* **2021**, *12*, 4831–4840.
14. Al Machot, F.; Elmachot, A.; Ali, M.; Al Machot, E.; Kyamakya, K. A deep-learning model for subject-independent human emotion recognition using electrodermal activity sensors. *Sensors* **2019**, *19*, 1659. [CrossRef]
15. Zhong, B.; Qin, Z.; Yang, S.; Chen, J.; Mudrick, N.; Taub, M.; Azevedo, R.; Lobaton, E. Emotion recognition with facial expressions and physiological signals. In Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, USA, 27 November–1 December 2017; pp. 1–8.
16. Dzedzickis, A.; Kaklauskas, A.; Bucinskas, V. Human Emotion Recognition: Review of Sensors and Methods. *Sensors* **2020**, *20*, 592. [CrossRef]
17. Raheel, A.; Majid, M.; Alnowami, M.; Anwar, S.M. Physiological Sensors Based Emotion Recognition While Experiencing Tactile Enhanced Multimedia. *Sensors* **2020**, *20*, 4037. [CrossRef]
18. Liu, S.; Wang, X.; Zhao, L.; Zhao, J.; Xin, Q.; Wang, S.-H. Subject-Independent Emotion Recognition of EEG Signals Based on Dynamic Empirical Convolutional Neural Network. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2020**, *18*, 1710–1721. [CrossRef] [PubMed]
19. Chao, H.; Liu, Y. Emotion Recognition From Multi-Channel EEG Signals by Exploiting the Deep Belief-Conditional Random Field Framework. *IEEE Access* **2020**, *8*, 33002–33012. [CrossRef]
20. Zheng, S.; Peng, C.; Fang, F.; Liu, X. A Novel Fuzzy Rough Nearest Neighbors Emotion Recognition Approach Based on Multimodal Wearable Biosensor Network. *J. Med. Imaging Heal. Inform.* **2020**, *10*, 710–717. [CrossRef]
21. Al Machot, F.; Ali, M.; Ranasinghe, S.; Mosa, A.H.; Kyandoghere, K. Improving subject-independent human emotion recognition using electrodermal activity sensors for active and assisted living. In Proceedings of the 11th Pervasive Technologies Related to Assistive Environments Conference, Corfu, Greece, 26–29 June 2018; pp. 222–228.
22. Santamaria-Granados, L.; Munoz-Organero, M.; Ramirez-Gonzalez, G.; Abdulhay, E.; Arunkumar, N. Using Deep Convolutional Neural Network for Emotion Detection on a Physiological Signals Dataset (AMIGOS). *IEEE Access* **2018**, *7*, 57–67. [CrossRef]
23. Rayatdoost, S.; Rudrauf, D.; Soleymani, M. Multimodal gated information fusion for emotion recognition from EEG signals and facial behaviors. In Proceedings of the 2020 International Conference on Multimodal Interaction, Online, 25–29 October 2020; pp. 655–659.
24. Siddharth, S.; Jung, T.-P.; Sejnowski, T.J. Utilizing Deep Learning Towards Multi-Modal Bio-Sensing and Vision-Based Affective Computing. *IEEE Trans. Affect. Comput.* **2022**, *13*, 96–107. [CrossRef]
25. Val-Calvo, M.; Álvarez-Sánchez, J.R.; Ferrández-Vicente, J.M.; Fernández, E. Affective robot story-telling human-robot interaction: Exploratory real-time emotion estimation analysis using facial expressions and physiological signals. *IEEE Access* **2020**, *8*, 134051–134066. [CrossRef]
26. Comas, J.; Aspandi, D.; Binefa, X. End-to-end facial and physiological model for affective computing and applications. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; pp. 93–100.
27. Huang, L.; Polanco, M.; Clee, T.E. Initial experiments on improving seismic data inversion with deep learning. In Proceedings of the 2018 New York Scientific Data Summit (NYSDS), New York, NY, USA, 6–8 August 2018; pp. 1–3.

28. Qin, F.; Gao, N.; Peng, Y.; Wu, Z.; Shen, S.; Grudtsin, A. Fine-grained leukocyte classification with deep residual learning for microscopic images. *Comput. Methods Programs Biomed.* **2018**, *162*, 243–252. [CrossRef]
29. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
30. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
31. Evmenenko, A.; Teixeira, D.S. The circumplex model of affect in physical activity contexts: A systematic review. *Int. J. Sport Exerc. Psychol.* **2022**, *20*, 168–201. [CrossRef]
32. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. AffectNet: A dataset for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **2017**, *10*, 18–31. [CrossRef]
33. Sharma, R.; Rajvaidya, H.; Pareek, P.; Thakkar, A. A comparative study of machine learning techniques for emotion recognition. In *Emerging Research in Computing, Information, Communication and Applications*; Springer: Singapore, 2019; pp. 459–464.
34. Kosti, R.; Alvarez, J.M.; Recasens, A.; Lapedriza, A. EMOTIC: Emotions in Context dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 61–69.
35. Song, S.; Jaiswal, S.; Sanchez, E.; Tzimiropoulos, G.; Shen, L.; Valstar, M. Self-supervised Learning of Person-specific Facial Dynamics for Automatic Personality Recognition. *IEEE Trans. Affect. Comput.* **2021**, *preprint*. [CrossRef]
36. Song, T.; Lu, G.; Yan, J. Emotion recognition based on physiological signals using convolution neural networks. In Proceedings of the 2020 12th International Conference on Machine Learning and Computing, Shenzhen, China, 15–17 February 2020; pp. 161–165.
37. Jeong, D.; Kim, B.-G.; Dong, S.-Y. Deep Joint Spatiotemporal Network (DJSTN) for Efficient Facial Expression Recognition. *Sensors* **2020**, *20*, 1936. [CrossRef] [PubMed]
38. Riaz, M.N.; Shen, Y.; Sohail, M.; Guo, M. eXnet: An Efficient Approach for Emotion Recognition in the Wild. *Sensors* **2020**, *20*, 1087. [CrossRef] [PubMed]
39. Patlar Akbulut, F. Hybrid deep convolutional model-based emotion recognition using multiple physiological signals. *Comput. Methods Biomech. Biomed. Eng.* **2022**, *online ahead of print*. [CrossRef]
40. Huang, Y.; Yang, J.; Liu, S.; Pan, J. Combining facial expressions and electroencephalography to enhance emotion recognition. *Future Internet* **2019**, *11*, 105. [CrossRef]
41. Bandyopadhyay, S.; Thakur, S.S.; Mandal, J.K. Online Recommendation System Using Human Facial Expression Based Emotion Detection: A Proposed Method. In *International Conference on Advanced Computing Applications*; Springer: Singapore, 2022; pp. 459–468.
42. Katsigiannis, S.; Ramzan, N. DREAMER: A dataset for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 98–107. [CrossRef] [PubMed]
43. Ahmed, M.M.; Khan, N.; Das, A.; Dadvar, S.E. Global lessons learned from naturalistic driving studies to advance traffic safety and operation research: A systematic review. *Accid. Anal. Prev.* **2022**, *167*, 106568. [CrossRef]
44. Swapna, M.; Viswanadhula, U.M.; Aluvalu, R.; Vardharajan, V.; Kotecha, K. Bio-Signals in Medical Applications and Challenges Using Artificial Intelligence. *J. Sens. Actuator Networks* **2022**, *11*, 17. [CrossRef]
45. Sciaraffa, N.; Di Flumeri, G.; Germano, D.; Giorgi, A.; Di Florio, A.; Borghini, G.; Vozzi, A.; Ronca, V.; Varga, R.; van Gasteren, M.; et al. Validation of a Light EEG-Based Measure for Real-Time Stress Monitoring during Realistic Driving. *Brain Sci.* **2022**, *12*, 304. [CrossRef]
46. Stoychev, S.; Gunes, H. The Effect of Model Compression on Fairness in Facial Expression Recognition. *arXiv* **2022**, arXiv:2201.01709.
47. Jia, X.; Zhou, Y.; Li, W.; Li, J.; Yin, B. Data-aware relation learning-based graph convolution neural network for facial action unit recognition. *Pattern Recognit. Lett.* **2022**, *155*, 100–106. [CrossRef]