Neural Computing and Applications (2021) 33:3155–3168
https://doi.org/10.1007/s00521-020-05209-7

ORIGINAL ARTICLE

# Deep CNN models-based ensemble approach to driver drowsiness detection

Mohit Dua[1] · Shakshi[1] · Ritu Singla[1] · Saumya Raj[1] · Arti Jangra[1]

Received: 29 January 2020 / Accepted: 13 July 2020 / Published online: 20 July 2020

## Abstract

Statistics have shown that many accidents occur due to drowsy condition of drivers. In a study conducted by National Sleep Foundation, it has been found that about 20% of drivers feel drowsy during driving. These statistics paint a very scary picture. This paper proposes a system for driver drowsiness detection, in which the architecture detects sleepiness of driver. The proposed architecture consists of four deep learning models: AlexNet, VGG-FaceNet, FlowImageNet and ResNet, which use RGB videos of drivers as input and help in detecting drowsiness. Also, these models consider four types of different features such as hand gestures, facial expressions, behavioral features and head movements for the implementation. The AlexNet model is used for various background and environmental changes like indoor, outdoor, day and night. VGG-FaceNet is used to extract facial characteristics like gender ethnicities. FlowImageNet is used for behavioral features and head gestures, and ResNet is used for hand gestures. Hand gestures detection provides a precise and accurate result. These models classify these features into four classes: non-drowsiness, drowsiness with eye blinking, yawning and nodding. The output of these models is provided to ensemble algorithm to obtain a final output by putting them through a SoftMax classifier that gives us a positive (drowsy) or negative answer. The accuracy obtained from this system came out to be 85%.

**Keywords** Deep learning · CNN · AlexNet · FlowImageNet · VGG-FaceNet · ResNet

## 1 Introduction

There have been many accidents and deaths due to drowsiness of the driver. Approximately 328,000 crashes occur every year in a country like the USA. These drowsy driving accidents cost the society $109 billion annually [1]. Many automobile companies are using various drowsy driver detection systems to ensure that their automobile is infallible. Companies like Audi, BMW, Bosch have some very effective and reliable drowsy detection systems like driver alert, driver attention warning systems. However, there is still a room for improvement.

In driver drowsiness detection systems, there are various types of features that can be used to detect drowsiness. Detection can be done by utilizing behavioral data, physiological measurements and vehicle-based data. Behavioral data comprise eye/face/head movement captured by using a camera. Physiological measurements comprise electrocardiogram (ECG) heart rate, electrooculogram (EOG), electroencephalogram(EEG), etc. [2]. Vehicle-based data are obtained from steering wheel motion, speed of vehicle, braking pattern and deviation obtained in a lane position. Measurements can be obtained through questionnaires and electrophysical measures. However, it is normally not possible or impractical to obtain reliable feedback from a driver in actual driving situation and all these techniques have their merits and demerits. Physiological

✉ Mohit Dua
er.mohitdua@nitkkr.ac.in

Shakshi
shakshi_11610269@nitkkr.ac.in

Ritu Singla
ritusingla1998@gmail.com

Saumya Raj
saumya.raj.sldav@gmail.com

Arti Jangra
artijangra28@gmail.com

[1] Department of Computer Engineering, National Institute of Technology, Kurukshetra, India

measurements are too intrusive, as these interfere with the ability of the driver to drive safely. Vehicle-based measurements require hardware and it may be too costly. Behavioral measurements on the other hand require minimal hardware, are very cost-effective and also, do not interfere with the driver's driving ability. All the merits of behavioral measurements have compelled us to use these as the base for the proposed detection system discussed in this paper. The work in this paper performs drowsiness detection by utilizing facial features, head movements, hand gestures and environmental features.

Frequent yawning, nodding or head swinging and constant blinking are considered as signs of drowsiness. Even the percentage of eyelids being closed over a particular time is a good measure to detect drowsiness and this parameter has been used in many commercial products. Also, various facial features like jaw dropping, inner brow and outer brow rise, lip movement for yawning, head motions are being used in recent researches for detection of drowsiness in drivers. In some of the earlier works, these features are measured handcrafted. But handcrafted measurements have lot of limitations especially, when the driver is wearing sunglasses and there is illumination variation. The detection becomes very vague in these real-world situations. So, to reduce this, people started to use machine learning methods to detect these facial features.

Ursulescu et al. detected drowsiness in eyes using eye blinks detection and measured the duration between every continuous eye blink [3]. Tabrizi et al. used image processing techniques by saturation or the S-channel of the color model to perform open or closed eye analysis for detecting drowsiness [4]. Dehnavi et al. also used image processing techniques for eye blink detection [5]. Rahman et al. proposed an architecture that uses eyes feature points to determine whether eye is open or closed [6]. The object detection was proposed by Viola–Jones by using cascade classifier, that had features for eye detection [7]. Lu et al. used texture features based on pixel-pattern and rectangle features to perform eye detection [8]. In the work by Rahul Atul Bhone, facial landmarks are used for detecting drowsiness. It detects eye using the Viola–Jones-based classifier and calculates the eyes aspects ratio (EAR) for every frame, which if below a threshold is an indication of a closed eye state [9].

The main objective of the system proposed by Lee et al. [10] was to detect the driver's drowsiness level on the basis of the driver's behavior derived from the motion data. These data were collected from the built-in motion sensors in the smartwatch, for example the accelerometer and the gyroscope. Eight features were selected that served as an input to a support vector machine (SVM) classifier. Zhenlong et al. [11] utilized the performance of K-nearest neighbor, SVM and artificial neural network (ANN) classifiers for driver drowsiness detection based on a driving simulator. In this, initially, vehicle performance measures were obtained through sensors, then six classifiers were built for six curve segments and one classifier for all straight segments. Finally, comparison of the performance of K-nearest neighbor, SVM and artificial ANN classifiers was done. The results showed that the SVM classifier had the fastest classification time and the highest accuracy. Driver drowsiness is a significant cause of fatal crashes every year in the world. Mahmoodi et al. [12] proposed an approach, where driver's drowsiness was determined by studying surface electromyography signal features. The electromyography signals measured were mid deltoid, clavicular portion of the pectoralis major, and triceps and biceps long heads from the upper arm and shoulder muscles. The five features range, variance, relative spectral power, kurtosis and shape factor were extracted. A binomial function was fitted for each feature and six classifiers were applied for classification. The approach proposed by Sadegh et al. [13] presented a novel feature selection technique which was used to design a noninvasive driver drowsiness detection system based on steering wheel data. The proposed attribute selector could determine the most related features to the drowsiness level to improve accuracy of the classification. This approach was based on the culmination of the filter and wrapper feature selection algorithms using adaptive neuro-fuzzy inference systems (ANFIS). The experimental data were collected from about 20.5 h of driving in the simulator. The results indicated that the drowsiness detection system works with a high accuracy. In the approach suggested by Gielen et al. [14], the use of a new set of features, for determining driver drowsiness based on physiological changes related to thermoregulation, was tested. Nineteen participants performed a driving simulation in which the temperature of the nose and wrist as well as the heart rate (HR) was monitored. On average, an initial enhancement in temperature, followed by a slow decrease, was obtained in drivers who experienced drowsiness. For non-drowsy drivers, no such trends were observed. A classification based on each of these variables resulted in a good accuracy, concluding that ultimately the state of 17 out of 19 drivers was detected correctly. Hence, it can be inferenced that the use of physiological features related to thermoregulation shows potential for future research in this field.

In the last few years, the deep learning-based methods came into research trends. Choi et al. in his paper used deep learning models to develop a gaze zone detection algorithm, which was based on convolutional neural networks (CNN) [15]. With the help of the CNN networks, the model learns features of the driver and these features are then fed to support vector machine (SVM) to detect the drowsiness. The work proposed by Dwivedi et al. in [16], used deep learning for detecting facial features. It used a 3-layered CNN for detecting facial features by the image of the driver, that are fed as input to the CNN network. In this 3-layer CNN, the output is extracted layer by layer and the output of final layer is the final output. And this output is then passed to a SoftMax classifier for classification as required. Another advancement of this work was proposed by Park et al. [17]. He used three features and used 3-CNN models to detect these three features. The features used were behavioral features, environmental and background, and facial features. The output of these models is classified into four categories as: non-drowsiness, drowsiness with yawning, nodding and eye blinking. The output of these models is fed to a SoftMax classifier and ensemble via two architectures, which they called as independently averaged architecture (IAA) and feature fused architecture (FFA). These architectures compare the three models by concatenating and integrating their fully connected (FC) seven layers, and based on this concatenation, the output is classified.

Review of the enhancements in the technologies used for driver drowsiness from hardware to machine learning and then, to deep learning, provides us a huge scope for further improvements. Motivated by the above discussed research works, the work proposed in this paper also uses deep learning approach and is an extension of the work done by Park et al. [17]. The proposed work provides solution by using 4 CNN models and assemble them using ensemble algorithm to get good results. Previously, the features used were behavioral, facial and environmental, which were detected by FlowImageNet, VGG-FaceNet, AlexNet, respectively. Here, the proposed work adds another set of features to detect yawning with hand gestures of the drivers while driving and applies ResNet model to classify these features.

All the four types of features, i.e., behavioral, environmental, facial and hand movements are detected by feeding the input as RGB video and optical flows of the driver while driving and are then separated into frames. The four deep learning-based models, i.e., FlowImageNet, AlexNet, VGG-FaceNet, ResNet are applied individually (one model on one set of features) to predict and classify the features. Finally, the output of four models is provided to ensemble algorithm to get final answer as drowsiness or non-drowsiness. Given an input sequence, the FlowImageNet model is trained using dense optical flow image, extracted from the frames of the video, as input. This training enables the model to learn the behavioral features such as facial expressions and head movement or nodding, which is the most basic indication of sleepiness. AlexNet has been used to handle the environmental features such as night time shade variation, effects of sunglasses and indoor/outdoor shade changes. VGG-FaceNet has been trained to measure all facial expressions such as lips, eye-pupil, eyebrows and chicks. VGG-FaceNet is also related to gender, originality and other accessories. As said earlier, all these models are independently trained and used for prediction. The output is classified into four classes: non-drowsiness, drowsiness with yawning, nodding and eye blinking. Finally, it is passed to ensemble algorithm to get a final output. The non-drowsiness category is symbolized as 0 and the other three categories are taken as 1. Simple averaging technique is used for ensemble. If the average of outputs from all models is greater than the threshold value 0.24, then the driver is said to be drowsy. The proposed architecture is trained on National Tsing Hua University (NTHU)-driver drowsiness standard video datasets [18] and the prediction results are shown in terms of detection accuracy. The proposed system has 85% accuracy and proves to be robust in every condition.

Rest of the paper is structured as: Sect. 2 describes the preliminaries, Sect. 3 discusses the proposed architecture, Sect. 4 shows and analyzes the results and Sect. 5 concludes the paper with discussion on future directions.

## 2 Preliminaries

### 2.1 Convolutional neural networks (CNN)

CNN, also called as ConvNet, is a deep learning algorithm, widely used in the image processing scenarios. The algorithm learns different weights and biases corresponding to various objects present in an input image, thereby differentiating images from one another. The algorithm takes an input image in the form of a matrix of pixel values, more precisely using its height, width, number of channels. The network comprises number of layers, the first of which is convolutional layer, where input image is fed. The convolutional layer comprises set of convolutional kernels, covering a small region of the image, which are convolved over the entire image to learn various features present in it. This convolution operation is expressed as:

$$Z_b^a = \left(X_{x,y} * C_b^a\right) \tag{1}$$

where $X_{x,y}$ represents input image, $x, y$ represents spatial locality and $C_b^a$ shows the $b$th convolutional kernel of $a$th layer. CNN comprises number of such convolutional layers along with activation and pooling layers. The activation layer contains an activation function, also known as decision function, which is responsible for analyzing nonlinear properties of an image. The pooling layers are responsible for image down sampling (i.e., reducing the volume of the image). The nonlinear layer follows each convolutional layer and the pooling layer follows each activation layer. The pooling operation is given as:

$$F_a = P_o\left(Z_{x,y}^a\right) \tag{2}$$

where $F_a$ represents $a$th output feature map, $Z_{x,y}^a$ represents $a$th input feature map and $P_o(.)$ represents the pooling operation type. CNN also contains a fully connected (FC) Layer, which follows a series of convolutional, nonlinear and pooling layers. The FC layer is attached at the end of the network and is responsible for producing an $N$-dimensional vector corresponding to $N$ number of output classes.

## 2.2 AlexNet

Earlier, CNN was supposed to be used for hand digit recognition tasks only and didn't scale well to all the classes of an image. To enhance the learning capacity of CNN, AlexNet model was introduced in 2012, which showed the solid results for image recognition and classification tasks. The AlexNet model made the CNN deeper by introducing a number of parameter optimization strategies. To make CNN applicable for diverse classification of images, the feature extraction stages were extended from 5 (in LeNet) to 7 (in AlexNet). However, the problems like overfitting and vanishing gradient descents arise, when depth of the model is increased. Overlapping subsampling and local response normalization are used to reduce overfitting and the large-sized filters (11 × 11 or 5 × 5) are introduced in initial layers.

## 2.3 VGG-FaceNet

VGG-FaceNet has very effective design principle. It is 16 layers deep as compared to previously proposed architectures like AlexNet. The small-sized filters, i.e., 3 × 3 sized are concurrently placed over the input matrix, inducing the effect of large-sized filters.

Because of the above features and homogeneous topology possessed by VGG-FaceNet, it is considered as the most preferable model for image localization and image classification. However, it suffers from the limitation of high computational cost because of the application of approximately 140 million parameters.

## 2.4 FlowImageNet

This model is 8 layers deep, comprising of 5 convolutional layers and 3 FC layers. The model is proposed to gain expertise in learning behavioral features of the entity present in an image. As in earlier proposed models, this model also includes ReLu activation function to resolve the overfitting and vanishing gradient descents issues.

## 2.5 ResNet

To overcome the issue of computational complexity suffered by previously proposed models, ResNet introduced the concept of residual learning. ResNet is 152 layers deep CNN, 8 and 20 times deeper than VGG-FaceNet and AlexNet, respectively, and has less computational complexity. Figure 1 shows the intuition behind using residual blocks in ResNet:

when the input and output dimensions are same, then the mathematical equation for output function becomes:

$$Z = F(x(\text{input}), \{w_k\}) + x(\text{input}), \tag{3}$$

and when the dimensions are different, then the mathematical equation for output function becomes:

$$Z = F(x(\text{input}), \{w_k\}) + w_s x(\text{input}) \tag{4}$$

where $w_k$ represents the weights associated with $k$th layer. It has been observed that ResNet with 50 layers or 101 layers or 152 layers produces very less error in image classification tasks. ResNet gives good performance on image recognition and localization tasks.

## 2.6 Ensemble

Neural networks are designed to recognize the patterns in data, i.e., their learning capability is made extremely high. However, along with all these merits of neural networks, there is a drawback that these networks are sensitive to the specifics of training data. Each time when they are trained, a different set of weights are obtained, thereupon producing different predictions. A better and successful approach, training multiple models instead of single model and obtaining final predictions by combining the results of each trained model, known as an ensemble approach, was proposed to resolve the issue associated with neural networks. There are various ensemble methods available such as weighted voting, simple averaging and GASEN [19]. In our solution, simple averaging approach is used. If the value is greater than the threshold value, then the driver is said to be

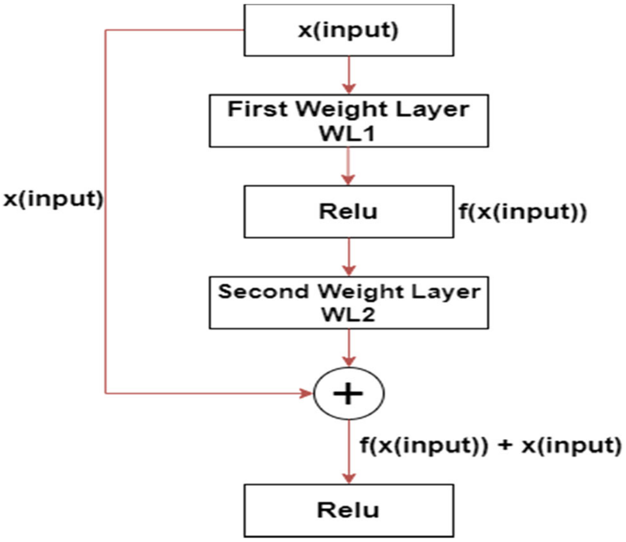


**Fig. 1** Residual block

drowsy. Undoubtedly, this method surpasses the single models prediction method.

## 3 Proposed architecture

This section describes the overall architecture of the proposed driver drowsiness detection system developed to work in various circumstances and real-world scenarios. This proposed architecture made up of two parts: learning feature representation and ensemble. For learning feature representation, four models such as FlowImageNet, AlexNet, VGG-FaceNet, ResNet are used. Each model is trained on National Tsing Hua University(NTHU)-driver drowsiness standard video datasets [18] to extract different features by controlling the network depth, i.e., number of convolution layers, number of FC layers, number of parameters and number of neurons in it. Table 1 shows the parameters used in training each network to learn or extract different features related to drowsiness.

After the model is trained or features are extracted from input image, each network is fine-tuned for multi-class drowsiness classification, which is performed by last SoftMax layer in each network. Ensemble strategy is used as if any of the model is showing positive result, i.e., showing drowsiness, which means there are some frames in the video which are showing drowsy behavior of the driver. It is to be kept in mind that the video of drowsiness contains some non-drowsiness part also which is reflected in frames, but a non-drowsy video will not contain any frames of drowsiness. The overall architecture of the Driver Drowsiness System is shown in Fig. 2 and the pseudocode is given in Algorithm 1. Here, video stream is given as input and frames are extracted from this input stream. These extracted frames are fed to the said deep learning models for training and prediction. All these four models work independently and give independent outputs. The models classify the videos into four categories as: non-drowsiness, drowsiness with yawning, nodding and eye blinking. These model outputs are ensemble and final output will be shown as: drowsiness or non-drowsiness. The following subsection discuss all the modules in detail.

**Algorithm 1:** Driver Drowsiness Detection System

**Input** : *Video*

**Code:**

```
DriverDrowsinessDetection(InputVideo)
{
    // Categorize the video according to Behavioral features
   FlowImageNetOutput=modelFlowImageNet(InputVideo)
// Categorize the video according to Environmental features
    AlexnetOutput=modelAlexnet(InputVideo)
// Categorize the video according to Facial feature representation
   VGG-16Output=modelVgg16(InputVideo)
   // Categorize the video according to hand gestures
   ResNetOutput=modelResNet(InputVideo)
   // Ensemble the outputs
   output=ensemble(AlexnetOutput,VGG-16Output,FlowImageNetOutput,ResNetOutput)
   if(output>threshold_value){
        print("Drowsy State")
   }
   else
   {
        print("NonDrowsy State")
   }
}
```

**Table 1** Different parameters values associated with each CNN network

| Model | No. of convolution layers | No. of FC layers | Number of parameters |
|---|---|---|---|
| AlexNet | 5 | 3 | 58,299,140 |
| VGG-FaceNet | 13 | 3 | 138,357,544 |
| FlowImageNet | 5 | 3 | 80,283,396 |
| ResNet | 48 | 2 | 23,587,712 |

### 3.1 Frames extraction

Videos are the collection of different frames arranged in a particular order. The videos cannot be fed directly to the models. For this reason, frames are extracted from the videos. The FPS (frames per second) rate used is 32, as lower values may result in overfitting. On the other hand, if we use higher values, drowsiness frames may get skipped. The work proposed in [17] used the FPS value 30; however, to accommodate the extra feature, the proposed work uses FPS value 32. OpenCV contains many different functions to perform multiple operations on videos. The step of frames extraction in the proposed architecture includes taking the video, breaking it into frames and saving the frames.

### 3.2 FlowImageNet for learning behavioral features

In the proposed architecture, model FlowImageNet is used to learn behavioral features such as movement patterns of face and head from the input video stream sequence. The proposed FlowImageNet is a pre-trained 8-layered CNN network having *conv1* to *conv5* as convolutional layers and *fc6* to *fc8* as fully connected layers [17]. The architecture and parameters used in proposed FlowImageNet are shown in Fig. 3.

A fine-tuned multi-classification training of the model is performed, and the image is down sampled to fixed resolution of 227 × 227 at the time of training. Training dataset is constructed by taking the optical dense flow images from the input video and passing them through the above discussed 8-layered CNN network to give the desired categorized output over the 4 classes.

### 3.3 AlexNet for learning environmental features

The AlexNet model is used to detect the drowsiness in various robust background and environmental conditions such as indoor/outdoor, day/night. It is a pre-trained model which is fed by frames extracted from the input video stream. It is a deep learning convolutional network which is made up of 8 layers and used far way better than other models for large scale image classification. Like FlowImageNet model architecture, it contains 5 convolutional and 3 fully connected layers and is also fine-tuned with multi-class classification [17]. The architecture of AlexNet used in the proposed work is shown in Fig. 4.

Here also, the input image is down sampled to the fixed resolution of 227 × 227. The first convolutional layer filters the input image of size 227 × 227 × 3 by 96 kernels of size 11 × 11 × 3 and strides of 4 pixels each. This is then passed through max pooling and normalization layers. The output of this is passed through second convolutional layer which has 256 kernels of size 5 × 5×48. Now in third, fourth and fifth layer, there are no max pooling and normalization. These layers are connected directly. Third convolutional layer has 384 kernels of size 3 × 3×256. The fourth convolutional layer consists of 384 kernels of size 3 × 3×192 and fifth convolutional layer consists of 256 kernels of size 3 × 3 × 192. The fully connected layers have 4096 neurons each. The output of last fully connected layer will be fed to 4-way SoftMax classifier to get the output.

### 3.4 VGG-FaceNet for learning facial features

VGG-FaceNet is used for learning facial features like gender and ethnicity from the input stream. VGG-FaceNet is a pre-trained 16-layer model and has been trained on various celebrity faces, YouTube dataset and also, on popular labeled faces of the world. As VGG-FaceNet has deeper structure, we used a shorter version, i.e., 8-layered VGG-FaceNet model. The layered architecture of VGG-FaceNet used by the proposed work is shown in Fig. 5. Although, it has a deeper structure, but here it has been described in a simpler way. Like the previously discussed two models, this model is also trained on fine-tuned multi-class classification. There are 4096 neurons in each fully connected layer and the output of *fc8* is passed to 4-way SoftMax classifier giving the desired categorized output over 4 classes.

### 3.5 ResNet for learning hand gestures

The ResNet model is used to give more precise results by detecting the yawning of the driver with hand on the mouth. Many times people yawn with the hand on mouth, but the facial features used earlier will not detect it as drowsiness. To overcome this, the proposed work uses an extra model for this feature only. The hand gestures along with facial features improve drowsiness detection. Also,

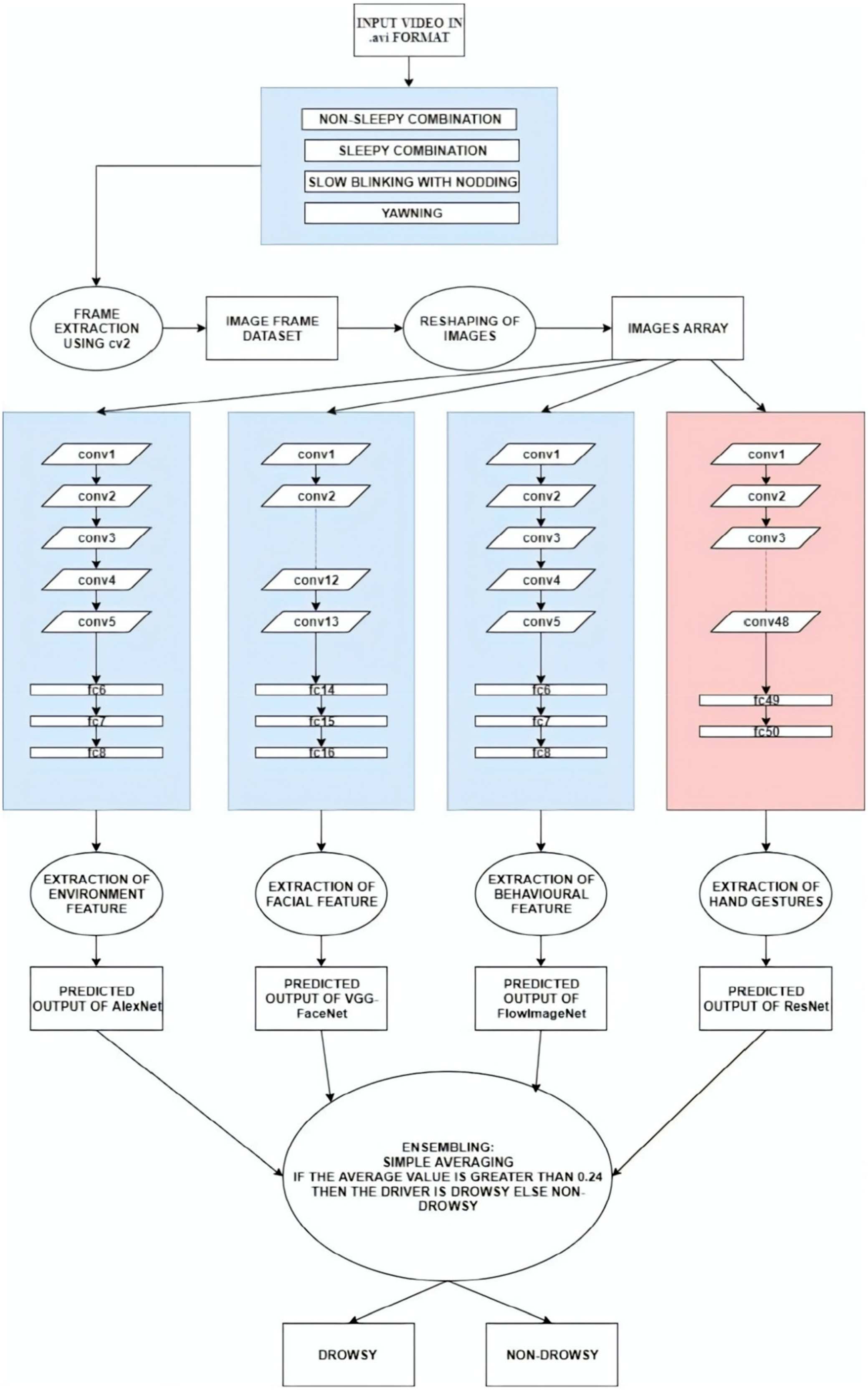

INPUT VIDEO IN .avi FORMAT
NON-SLEEPY COMBINATION
SLEEPY COMBINATION
SLOW BLINKING WITH NODDING
YAWNING
FRAME EXTRACTION USING cv2
IMAGE FRAME DATASET
RESHAPING OF IMAGES
IMAGES ARRAY
conv1
conv2
conv3
conv4
conv5
fc6
fc7
fc8
conv1
conv2
conv12
conv13
fc14
fc15
fc16
conv1
conv2
conv3
conv4
conv5
fc6
fc7
fc8
conv1
conv2
conv3
conv48
fc49
fc50
EXTRACTION OF ENVIRONMENT FEATURE
EXTRACTION OF FACIAL FEATURE
EXTRACTION OF BEHAVIOURAL FEATURE
EXTRACTION OF HAND GESTURES
PREDICTED OUTPUT OF AlexNet
PREDICTED OUTPUT OF VGG-FaceNet
PREDICTED OUTPUT OF FlowImageNet
PREDICTED OUTPUT OF ResNet
ENSEMBLING: SIMPLE AVERAGING IF THE AVERAGE VALUE IS GREATER THAN 0.24 THEN THE DRIVER IS DROWSY ELSE NON-DROWSY
DROWSY
NON-DROWSY

**Fig. 2** Proposed architecture

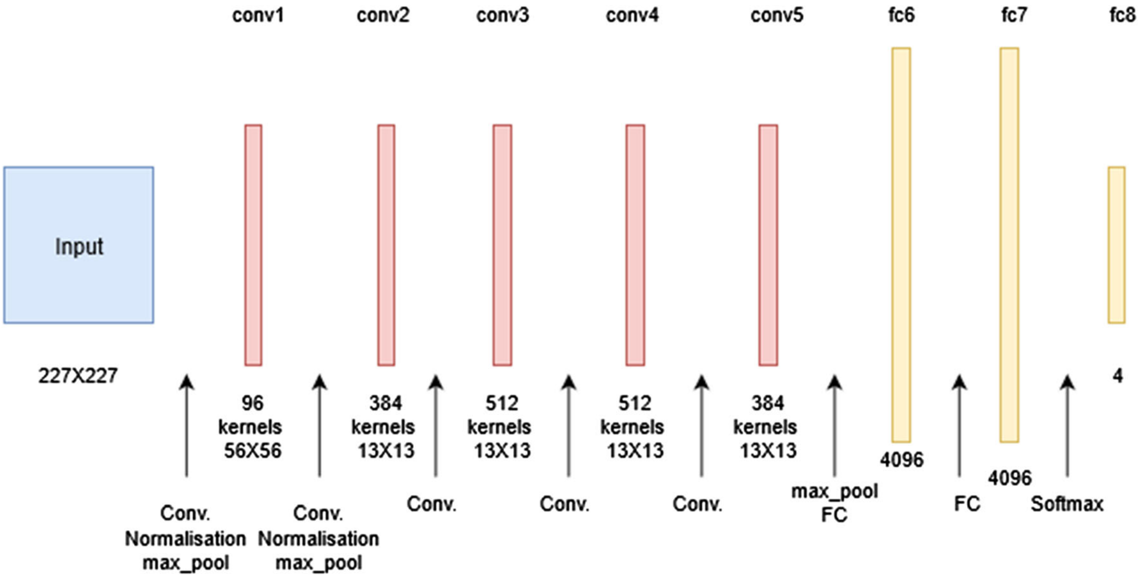


**Fig. 3** FlowImageNet model

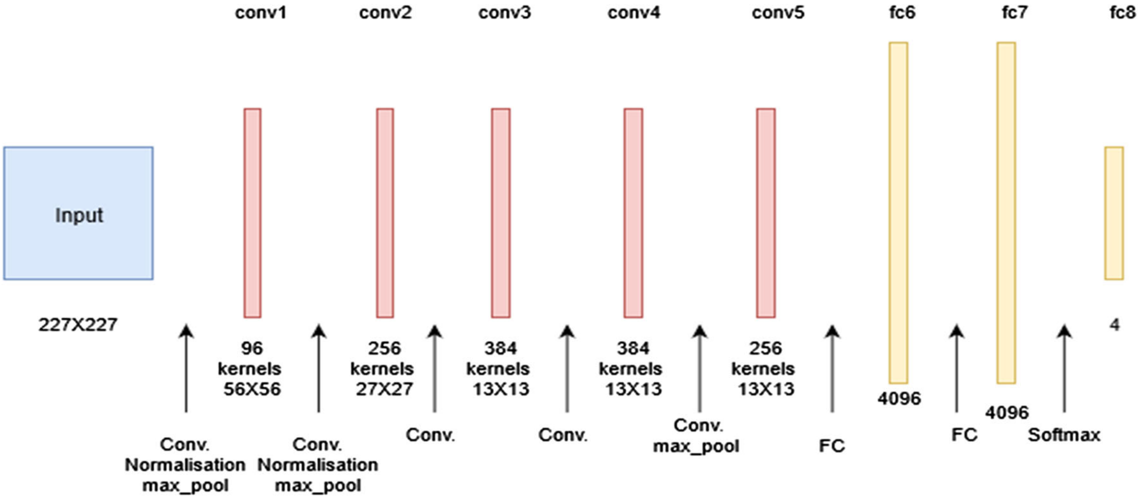


**Fig. 4** AlexNet model

use of bottleneck architecture by ResNet reduces the complexity of the proposed work.

The ResNet in the proposed architecture, as shown in Fig. 6, consists of different convolutional layers in which pooling, normalization and ReLu activation function are performed. These convolutional layers are followed by fully connected layer *fc* (84). The result of this fully connected layer is passed to a SoftMax classifier that gives the desired categorized output over 4 classes.

### 3.6 Ensemble algorithm

Ensemble algorithm increases the accuracy of the result. The individual output of all models is assembled using the technique of simple averaging to give a single output. In the proposed ensemble architecture, the output of each model is taken and averaged. A threshold value is set between the non-drowsy and drowsy state. If the result is larger than the threshold value, then the driver is said to be drowsy, else non-drowsy. The threshold value of 0.24 gives the best accuracy of 85%.

## 4 Simulation results

This section describes the experimental setup, the experiments conducted and results obtained while implementing the proposed driver drowsiness detection system. Also, it

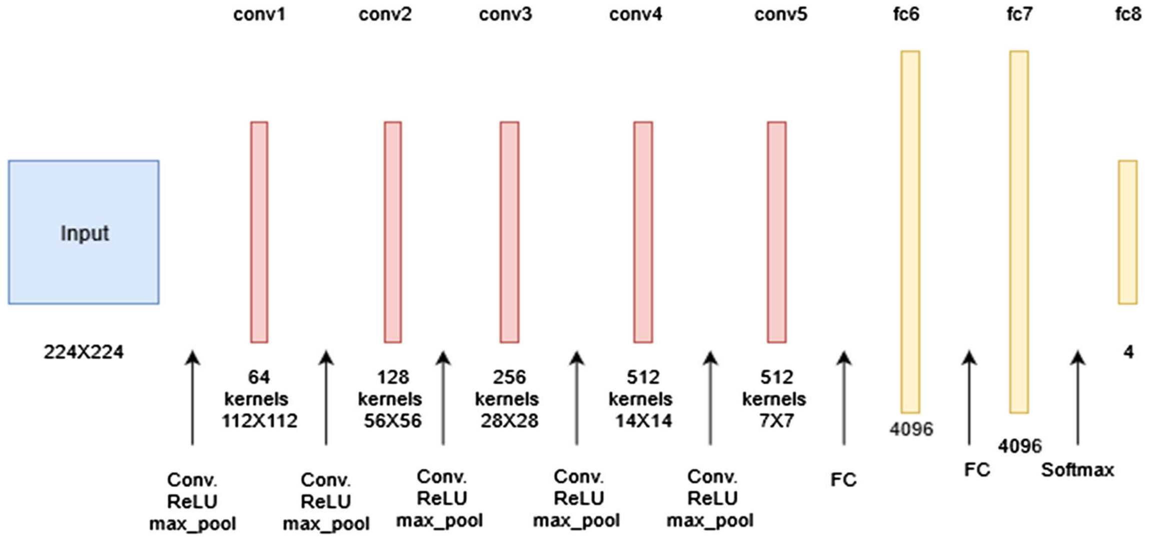


**Fig. 5** VGG-FaceNet model

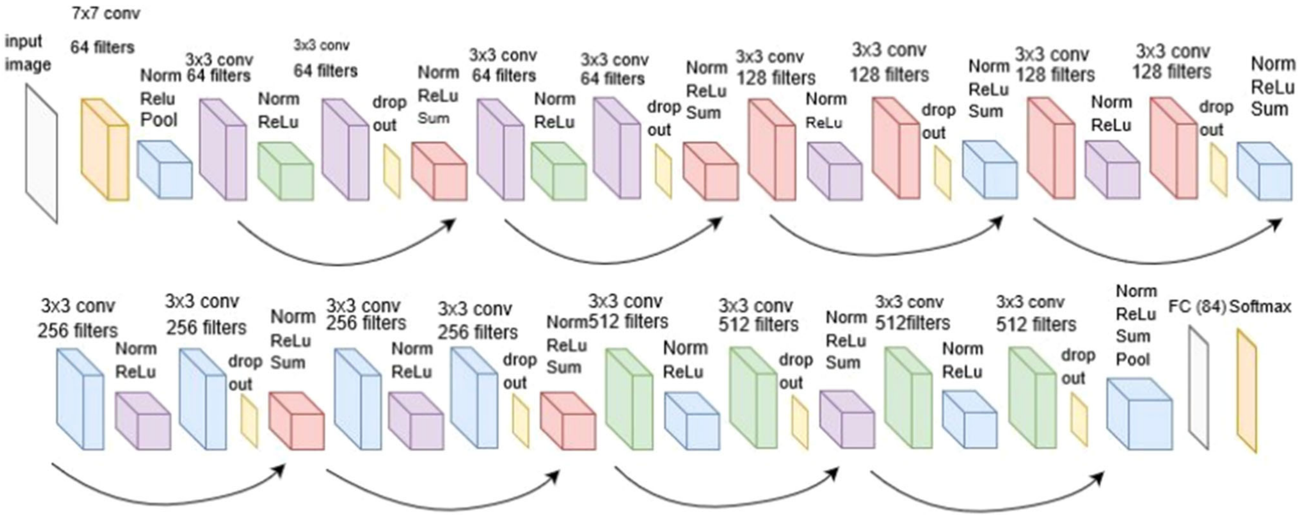


**Fig. 6** ResNet model [20]

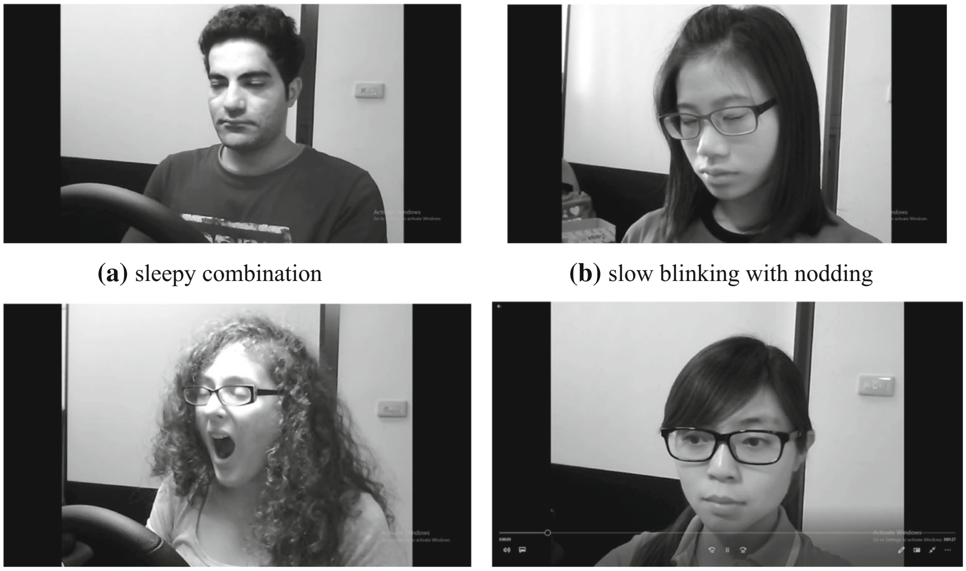
gives a comparison with previous approaches. The results also analyze the 4 CNN models in various situations. Each of these 4 CNN models provides favorable results on the dataset comprising video files. The best results are gained by using ensemble approach, as it uses a combination of the results given by each model trained separately.

### 4.1 Experimental setup

The proposed work has been implemented by using Python 3.6 on Jupyter Notebook platform and Windows 10 based PC that has Intel Core i5 processor with 2.4 GHz CPU along with 8 GB RAM. Videos of 640 × 480 pixels, 30 frames per second AVI format without audio have been used for conducting the experiments. Param Shavak [21] supercomputer has been used for training the proposed system model and MobaXterm has been used for accessing the same.

### 4.2 Video dataset

The NTH Drowsy Driver Detection (NTHU-DDD) video dataset [18] has been used for the implementation of the proposed system. It comprises people of different genders,



**(a)** sleepy combination

**(b)** slow blinking with nodding

**(c)** yawning

**(d)** non-sleepy combination

**Fig. 7** **a–d** Example frames of the training dataset [18]

**Fig. 8** Example frames of the testing dataset

ethnicities and ensures that the proposed system remains efficient and robust in all circumstances. Two hundred video clips of 25 subjects have been taken as training datasets and 50 video clips of six subjects are taken as testing datasets. Figure 7a–d shows some example frames from the video dataset of the proposed system taken from (NTHU-DDD) [18]. During training and evaluation, every frame is labeled by using the two markers, i.e., either drowsy or non-drowsy. Accordingly, output will be given in the form of two outcomes, i.e., either drowsy or non-drowsy obtained from the earlier defined drowsy features such as blinking, nodding, yawning and hand gesture positions.

The above mentioned dataset has been used for training, whereas the proposed approach has been tested on our own created test dataset and evaluated with live data. Figure 8 shows the example frame used for the test dataset. Each convolutional neural network (AlexNet, VGG-FaceNet, FlowImageNet, ResNet) of the approach was trained separately using the NTHU dataset [18].

### 4.3 Experiments and result analysis

The proposed approach aims to classify each frame in the videos based on feature extraction and learning with the help of different CNN models. Data augmentation technique has been used to enhance the durability and diversity of the implemented approach. After adjusting the pixel intensity of each of the images by 10%, 20%, −10%, −20% and normalization within the range [0–255], these training images are delivered into the multilayer CNN models. These networks have been trained using fine tuning and by modifying last fully connected and SoftMax layers. The number of output classes from the last layer comes out to be 4. The weights have been learned using backpropagation, which are used as learned feature representation. Convolving these weights with input images provides us the final feature representation that classifies drowsiness or non-drowsiness. In the proposed approach, the output of *fc7* layer is treated as trained feature representations. Since the classifier has been trained, the evaluation is done with the evaluation dataset and it was tested on the trained classifier. The final accuracy after training, testing and ensemble came out to be 85%. These favorable results are encouraging that deep learning is an efficient, intuitive and cost-effective method for market-wide adoption in driver drowsiness systems. This field continues to evolve and attract more researchers thus improving the results in future.

ResNet is a deep learning CNN model most commonly used for image classification. This model has above 23 million trainable parameters, that ensures a deep architecture which makes it better for image recognition. There are many more pre-trained deep models to use for example AlexNet, GoogleNet or VGG19, but the ResNet-50 has excellent generalization performance with lesser error rates on recognition tasks and is therefore more effective in enhancing the accuracy [22, 23]. Another important reason for enhancement of accuracy is the additional feature added in the proposed work. The proposed work adds hand gestures while yawning as an extra factor for considering drowsiness. Many times people yawn with their hand on mouth. The hand gestures along with facial features improve drowsiness detection. Figure 9a–d describes the deep learning models' accuracy evolution during training and validation while learning their respective features for driver drowsiness detection.

Table 2 demonstrates the accuracies of all the models for four different features. It also shows the final accuracy after ensemble of these models.

Table 3 gives the confusion matrix for the proposed work, where confusion matrix is defined as a table that is used to describe the performance of classification model on the basis of test data. In the proposed work, True Positives (TP = 82) denotes the correct prediction that the person is drowsy, whereas True Negatives (TN = 87) denotes the correct prediction that the person is not drowsy. Similarly, False Positives (FP = 13) and False Negatives (FN = 18) denote incorrect prediction the person is drowsy and the person is not drowsy, respectively.

Equations (5) to (8) describe the mathematical equations of various metrics that provide valuable information about the efficiency of the proposed model. The sensitivity or recall described by Eq. (5) defines how often the prediction is correct, when the actual value is positive. It is also known as "True Positive Rate" or "Recall." The term specificity given by Eq. (6) defines how often the prediction is correct when the actual value is negative. Precision metric is the number of values positive class predictions that actually belongs to positive class, which is given by Eq. (7). F1 score, given by Eq. (8), is the score that balances the concerns of both precision and recall in a single number.

$$\text{Sensitivity} = \frac{TP}{FP + FN} = 82/100 = 0.82 \tag{5}$$

$$\text{Specificity} = \frac{TN}{TN + FP} = 87/100 = 0.87 \tag{6}$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{82}{82 + 13} = 0.863 \tag{7}$$

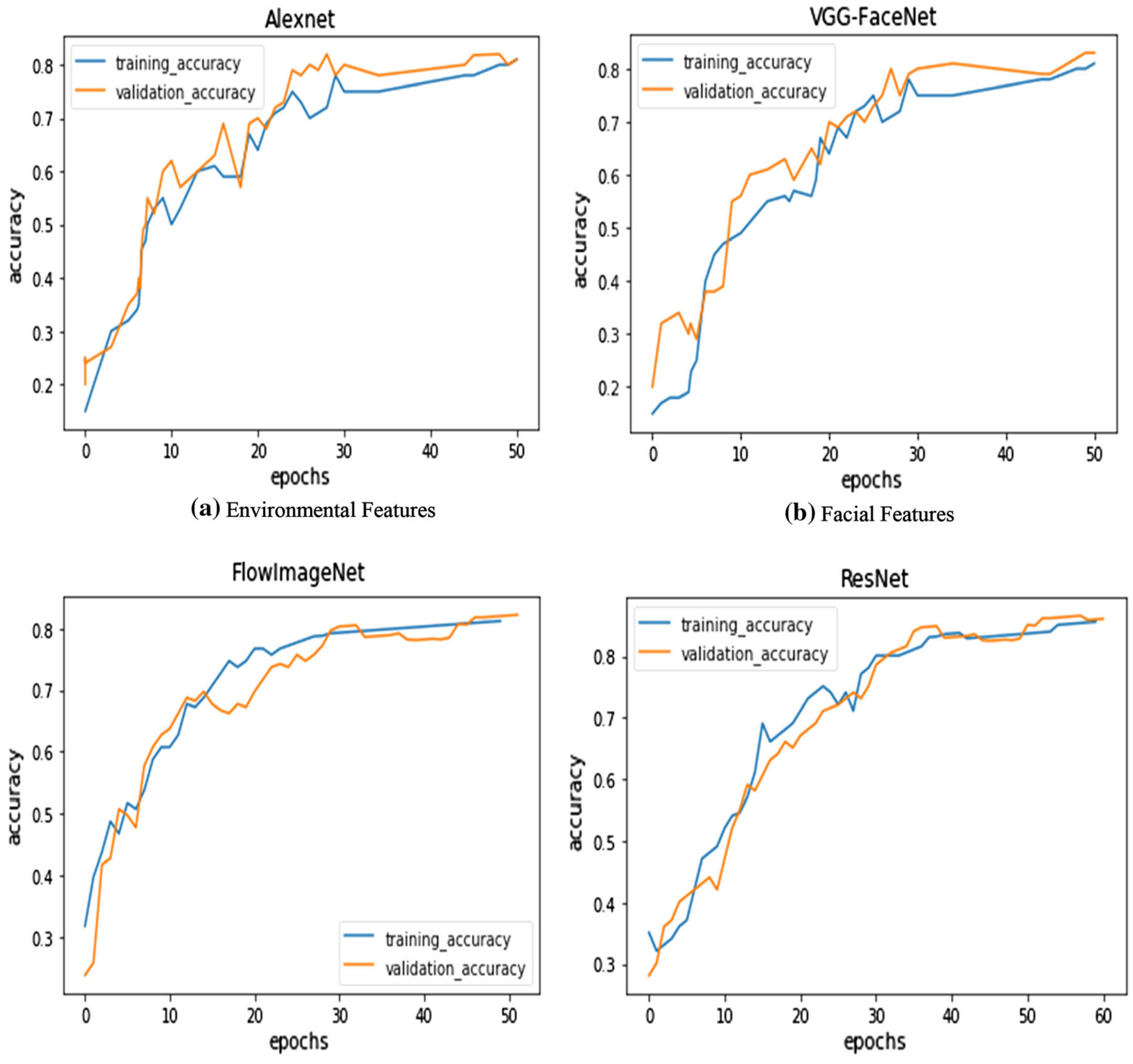
$$F1\text{Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = 2 * \frac{0.863 * 0.82}{0.863 + 0.82} = 0.8409. \tag{8}$$

### 4.4 Analysis and comparison with earlier proposed approaches

The domain for DDD has been researched extensively using machine learning, vehicular and electrophysiological techniques, but not much has been done by utilizing deep learning models. A very few researchers are working on it and one such comparison with a previous benchmark performance on this dataset has been provided. Table 4 demonstrates the comparison of the proposed work with an approach provided by students of KAIST university.

## 5 Conclusion and future work

The proposed deep learning model learns different features of a driver from the dataset which comprises images of driver in different states to detect the drowsiness and non-drowsiness state. Earlier proposed approaches focused on very few features like eye blinking, facial expressions, etc., as a result of which the predictions and accuracy were not completely up to the mark. The model suggested in this experimentation considers new features as well, such as head movements and hand gestures, thereby gives the most promising results. Furthermore, ensemble is used which ensures that the complementary nature of used models is fully utilized giving a more reliable output. This experimentation is based on the behavioral features of a driver only, vehicle and physiological measurements are not considered because they are too pricey, intrusive and not completely efficient. Physiological measurements are intrusive and the sensors and equipment are very expensive, so in the future if due to technology advances in the hardware field, these equipment become less costly and meddlesome, then it would be possible to incorporate these



**(a)** Environmental Features

**(b)** Facial Features

**(c)** Behavioural Features

**(d)** Yawning with hand gestures.

**Fig. 9 (a–d)** Accuracy evolution during training and validation

**Table 2** Drowsiness detection accuracies (%)

| Type of features | AlexNet | VGG-FaceNet | FlowImageNet | ResNet | Ensemble |
|---|---|---|---|---|---|
| Facial | 76.48 | 87.09 | 83.11 | 81.34 | 85% |
| Environmental | 85.94 | 79.19 | 78.12 | 84.89 | |
| Behavioral | 82.09 | 75.11 | 86.14 | 77.12 | |
| Yawning with hand gestures | 78.11 | 79.66 | 83.11 | 87.05 | |

**Table 3** Confusion matrix

| | (Actual) positive | (Actual) negative |
|---|---|---|
| (Predicted) positive | (TP) 82 | (FP) 13 |
| (Predicted) negative | (FN) 18 | (TN) 87 |

with the behavioral data to obtain a more reliable and complementary result. Furthermore, using complex deep learning models on more diverse dataset would also extract additional features, thereby providing a more satisfying result.

**Table 4** Analysis and comparison with earlier proposed approaches

| Parameters | Park et al. [17] | Nabit A. et al. [24] | Proposed approach |
|---|---|---|---|
| Number of models used | 3 | 2 | 4 |
| Models | AlexNet, VGG-FaceNet, FlowImageNet | Mobilenet CNN, SSD (Single Shot MultiBox Detector) | AlexNet, VGG-FaceNet, FlowImageNet, ResNet |
| Features | Behavioral, facial, environmental | Behavioral | Behavioral, facial, environmental, yawning with hand gestures |
| Model parameters | 1. AlexNet-60,000,000<br>2. VGG-FaceNet-60,000,000<br>3. FlowImageNet-60,000,000 | – | 1. AlexNet-58,299,140<br>2. VGG-FaceNet-138,357,544<br>3. FlowImageNet-80,283,396<br>4. ResNet-23,587,712 |
| Accuracy | 73.06% | 83.7% | 85% |
| Observations | Extracting environmental features from AlexNet, facial features from VGG-FaceNet and behavioral features from FlowImageNet and then ensemble the result using independently averaged architecture and feature fused architecture | Incoming video is passed to Mobilenets CNN and SSD, if the counter value is more than threshold, then it is defined as drowsy | Extracting environmental features from AlexNet, facial features from VGG-FaceNet, behavioral features from FlowImageNet and yawning with hand gestures features from ResNet and assembling the results manually |

## Compliance with ethical standards

**Conflict of interest** The manuscript does not have any conflict of interest.

## References

1. National Safety Council (2020) Drivers are falling asleep behind the wheel https://www.nsc.org/road-safety/safety-topics/fatigued-driving
2. Grace R, Byrne VE, Bierman DM, Legrand JM, Gricourt D, Davis BK, Staszewski JJ, Carnahan B (1998) A drowsy driver detection system for heavy vehicles. In: 17th DASC. AIAA/IEEE/SAE. Digital avionics systems conference. Proceedings (Cat. No. 98CH36267) vol 2, pp I36–1. IEEE
3. Ursulescu O, Ilie B, Simion G (2018). Driver drowsiness detection based on eye analysis. In: 2018 International symposium on electronics and telecommunications (ISETC) pp 1–4. IEEE
4. Tabrizi PR, Zoroofi RA (2008). Open/closed eye analysis for drowsiness detection. In: 2008 First workshops on image processing theory, tools and applications pp 1–7. IEEE
5. Dehnavi M, Attarzadeh N, Eshghi M (2011) Real time eye state recognition. In: 2011 19th Iranian conference on electrical engineering pp. 1–4. IEEE
6. Rahman A, Sirshar M, Khan A (2015) Real time drowsiness detection using eye blink monitoring. In: 2015 National software engineering conference (NSEC) (pp 1–7). IEEE
7. Viola P, Jones M (2001). Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, vol. 1, pp I–I. IEEE
8. Lu H, Zhang W, Yang D (2007) Eye detection based on rectangle features and pixel-pattern-based texture features. In: 2007 International symposium on intelligent signal processing and communication systems. pp 746–749. IEEE
9. Bhone RA (2019). Computer vision based drowsiness detection for motorized vehicles with web push notifications. In: 2019 4th International conference on internet of things: smart innovation and usages (IoT-SIU). pp 1–4. IEEE
10. Lee Boon-Leng, Lee Boon-Giin, Chung Wan-Young (2016) Standalone wearable driver drowsiness detection system in a smartwatch. IEEE Sens J 16(13):5444–5451
11. Li Zhenlong, Zhang Qingzhou, Zhao Xiaohua (2017) Performance analysis of K-nearest neighbor, support vector machine, and artificial neural network classifiers for driver drowsiness detection with different road geometries. Int J Distrib Sens Netw 13(9):1550147717733391
12. Mahmoodi Mohammad, Nahvi Ali (2019) Driver drowsiness detection based on classification of surface electromyography features in a driving simulator. Proc Inst Mech Engineers Part H J Eng Med 233(4):395–406
13. Arefnezhad S et al (2019) Driver drowsiness detection based on steering wheel data applying adaptive neurofuzzy feature selection. Sensors 19(4):943
14. Gielen Jasper, Aerts Jean-Marie (2019) Feature extraction and evaluation for driver drowsiness detection based on thermoregulation. Appl Sci 9(17):3555
15. Choi IH, Hong SK, Kim YG (2016) Real-time categorization of driver's gaze zone using the deep learning techniques. In: 2016 International conference on big data and smart computing (BigComp), pp. 143–148. IEEE
16. Dwivedi K, Biswaranjan K, Sethi A (2014) Drowsy driver detection using representation learning. In: 2014 IEEE International advance computing conference (IACC) pp 995–999. IEEE
17. Park S, Pan F, Kang S, Yoo CD (2016) Driver drowsiness detection system based on feature representation learning using various deep networks. In: Asian conference on computer vision, Springer, Cham. pp 154–164
18. Weng CH, Lai YH, Lai SH (2016) Driver drowsiness detection via a hierarchical temporal deep belief network. In: Asian conference on computer vision Springer, Cham, pp 117–133
19. Zhou ZH, Wu J, Tang W (2002) Ensembling neural networks: many could be better than all. Artif Intell 137(1–2):239–263

20. Alif MAR, Ahmed S, Hasan MA (2017) Isolated Bangla handwritten character recognition with convolutional neural network. In: 2017 20th International conference of computer and information technology (ICCIT), pp 1–6. IEEE
21. Agrawal S, Das S, Valmiki M, Wandhekar S, Moona R (2017) A case for PARAM Shavak: ready-to-use and affordable supercomputing solution. In: 2017 International conference on high performance computing & simulation (HPCS), pp 396–401. IEEE
22. Tian X, Chen C (2019) Modulation pattern recognition based on Resnet50 neural network. In: IEEE 2nd International conference on information communication and signal processing (ICICSP), Weihai, China, 2019, pp 34–38, https://doi.org/10.1109/icicsp48821.2019.8958555
23. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
24. Shakeel MF, Bajwa NA, Anwaar AM, Sohail A, Khan A (2019) Detecting driver drowsiness in real time through deep learning based object detection. In: International work-conference on artificial neural networks, Springer, Cham pp 283–296