A Monitoring System for Multiple Distracted States of Driver using Deep Learning

Asha T[1], Mahaveer[2], Manish J[3], Mohammed Farman[4], Monish J Hallegere[5]

[1]Professor, Dept. of Computer Science and Engineering, Bangalore Institute of Technology,Karnataka, India
[2,3,4,5]Dept. of Computer Science and Engineering, Bangalore Institute of Technology, Bengaluru, India

Abstract: *The choices and actions of the driver have a significant impact on the safety of the driver on the road. Understanding intelligent transportation systems, human-vehicle systems, and intelligent vehicle systems requires modelling and identifying driver behaviour.Most of the accidents are caused due to distracted state of the driver. In this paper, we are considering ten distracted states of driver i.e. looking at the mirror, drinking, operating radio, using phone, talking with co- passenger, talking with back seat passenger, talking on a phone, etc. We have used state farm dataset for training and testing the model. A driver activity recognition system is created using deep convolutional neural networks (CNN) in order to comprehend the driver's behaviour . As a result of distracted state, the model provides an alarm to alert the driver and a mail to concerned driver or an organization is sent. ResNet50 is used for identification of driver behaviour and results produce an accuracy of 98.87.*

***Keywords:* Deep learning, Convolutional neural network (CNN), Intelligent transport system, Distracted states.**

1. Introduction

According to the Indian government's statistics, there were 3,74,397 unintentional deaths in India in 2020, with vehicle crashes accounting for more than 35% of these deaths. According to the National Crime Records Bureau's (NCRB) annual report, there were 4,21,104 unintentional deaths in 2019 . According to the NCRB data, there were 5,88,738 such cases documented in 2020, with 3,38,903 people suffering injuries. The causes include human error, either intentionally or negligently.Here deliberate or negligent conduct includes getting bored or distracted from driving. Any activity that prevents the driver from paying attention to the road is considered a kind of distracted driving. Acts such as drinking, eating, doing makeup etc. While driving for longer distances, people inevitably get bored and occupy themselves with these tasks. This not only endangers the driver, but also the passengers and the occupants of other vehicles with which it crashes.

There are two ways in which we can broadly categorize these avoidance methods. One way is to employ autonomous or semi-autonomous cars to navigate the roads. The other is to keep the driver from getting distracted.

Autonomous and semi-autonomous vehicles are still under performance review and lot of research is being done on them to make them a commercial success and avoid any untoward accidents [1], customers also have to buy new cars to avail this technology and a study conducted [2] indicated that drivers fail to monitor the automation and detect critical signals due to over trust in these systems.

A more feasible solution is to retrofit the vehicle with technologies which aid in maintaining the driver's attention on the road ahead. Images of several distracted states are taken as an input using a camera mounted on the vehicle. The input is segmented using Gaussian Mixture Modeling (GMM) algorithm and the CNN model classifies the distracted states [3]. Some methods only use simple CNN and Recurrent neural networks (RNNs) to classify the states but the real time processing of the live data is a real hassle as processing delays are inevitable [4]. Instances are available where CNN is itself modified to increase accuracy, decrease computational time.

Decreasing-HOG convolutional neural network (D-HCNN) is such an instance. Here, HOG (Histogram of Gradients) features are used as an input. This eliminates useless background information while extracting the action characteristics of the task [6].

Systems using a single stream of images as an input are at a disadvantage [3][4] because of the blind-spots. These blind-spots cause vastly different gestures such as eating or drinking while talking on the mobile to be put in a same category. Methods aiming to curb this whereas other methods use multi-stream inputs. These inputs require robustly stabilized camera systems with high quality image streams [5].

Researchers have also proposed to use the architectures like GoogleNet, ALexNet, DenseNet, MobileNet, InceptionV3 and VGG-16 to classify the distracted states [8].
The models are also prone to over-fitting [7], there is also a possibility of data leakage between test and training data set [8]. Not only cameras, but sensors too areused for detecting reckless behavior. Systems that use accelerometers and gyro-sensors built into smartphones can identify this issue when the car is used in a remote location or on uneven, mountainous terrain. [9][10].

2. Related Works

In studies on human factors and intelligent transportation systems, driver behaviour analysis is a common subject. Researchers have developed a number of approaches to examine important visual data for identifying distracted driving using image and video data in recent years.These researches were mainly based on head parameter i.e. drowsiness, yawning, etc. Yang Xing et al. [1] implemented a system to detect driveris distracted or not. Initially, the Kinect camera is used to gather raw RGB photos. The GMM method is then used to segment the cropped images. Finally, for the activity's recognition task, the CNN model is used. Monagih Alkinani et al. [2] proposed model which basically looked upon into two major categories: Distraction and Fatigue/Drowsiness. Deep learning-based systems such as CNN, RNN was used. Chaoyun Zhang et al. [3] developed an original driver behavior recognition system that can classify user actions accurately in real time using input from in-vehicle cameras. A brand-new CNN architecture known as the D- HCNN was presented by Binbin Qin et al. [4]. It uses HOG features as input data, which can remove unnecessary background information and also suggest a new network based on reducing the filter size structure that can successfully extract the action characteristics of a task while significantly reducing the number of network parameters.

By merging three advanced deep learning models into one- the residual network (ResNet), the hierarchical recurrent neural network (HRNN), and the Inception , Munif Alotaibi et al. [5] proposed an architecture. As a component of the framework for recognising human actions, this study examines the posture recognition of distracted drivers. A model that Yang Xing et al. [6] presented is intended to identify seven activities carried out by various drivers. Normal driving, checking the right and left mirrors, and normal driving (front gazing) are the four tasks that fall under this normal driving category.The feature selection and extraction methods are constructed based on RFs and the MIC technique.

A real-time distracted driving detection system was proposed by Duy Tran et al. [7] and is achieved by using four deep CNN architectures, including VGG-16, AlexNet, and GoogleNet. The four networks' performances are assessed and contrasted in terms of accuracy and real-timeness to determine which algorithm is most appropriate for real-time deployment.

By taking into account the numerous driver characteristics, such as the driver's eyes, lips, blinking eyes, and head posture, Omar Wathiq et al. [8] established the effectively automated framework for driver distraction detection. Based on HOG feature extraction techniques, this framework is used with a set of publicly accessible driver tiredness data that includes videos of various drivers with various levels of exhaustion.

Ankit Pala et al. [9] proposed a model that deals with the problem of manual distractions and ways to identify and tackle them using different pre-trained models like InceptionV3, DenseNet, and MobileNet. The network is implemented based on the two data sources, one being the input image collected from the camera, other being the Inertial Measurement Unit (IMU) collected from the mobile. When splitting the data into training and testing data using the basic splitting technique of 80% in training and 20% in testing, there is leakage in data.

Using deep learning, Hesham M. Eraqi et al. [10] created a model to detect and identify distraction. A camera situated above the dashboard produces RGB images. Multiple convolutional neural network topologies can be trained and benchmarked using these images.

According to the related works mentioned above, methods were made suitable for drowsiness detection or considered only few distracted states. It was also found that the methods were analyzed in the low-level and lesser information was provided to the system which did not produce more promising result.

3. Methodologies

3.1 Deep learning

Deep learning is a machine learning technique that learns features and tasks directly from the data. The inputs are run through the neural network, where these neural networks have hidden layers. Deep learning has recently demonstrated outstanding performance and has dominated visual identification tasks. Particularly with regard to image identification tasks, the convolutional neural network (CNN) deep learning algorithm has made great progress. Finding the ideal CNN architecture is still a very difficult endeavour, though. As a result, numerous architectures have been suggested in the past, including GoogleNet (also known as Inception), AlexNet, VGGNet, and most recently, the deep residual network (also known as ResNet), which is now employed in our suggested system.

It has been demonstrated that Deep Convolutional neural networks are excellent at distinguishing various features from the images, and that stacking additional layers generally improves accuracy. The question of whether model performance can improve as we keep adding layers to the model arises. These inquiries lead to the issue of vanishing/exploding gradients. These issues were mainly resolved in a variety of methods, allowing networks with tens of layers to converge. However, as deep neural networks begin to converge, another issue arises: the accuracy quickly degrades due to saturation. In contrast to what we may expect, adding more layers to the appropriate deep learning model only raised the training error. This was not driven by overfitting as we assume.

3.2 CNN

When it comes to identifying patterns in the input image, such as lines, gradients, circles, or even eyes and faces, CNN performs well. Convolutional neural networks are extremely effective for computer vision because of this feature. Each of the numerous convolutional layers that make up CNN is capable of identifying more complex shapes. These layers are stacked on top of one another. The Input layer, Convolutional layer, Pooling layer, Fully Connected layer, and Output layer are all depicted in Fig. 1 in the CNN architecture.

Handwritten digits can be recognised with three or four convolutional layers, while human faces can be distinguished with 25 layers. The goal is to train robots to see the environment similarly way humans do, and to use this understanding for a variety of tasks like picture and video recognition, image inspection and categorization, media reconstruction, recommendation systems, natural language processing, etc.

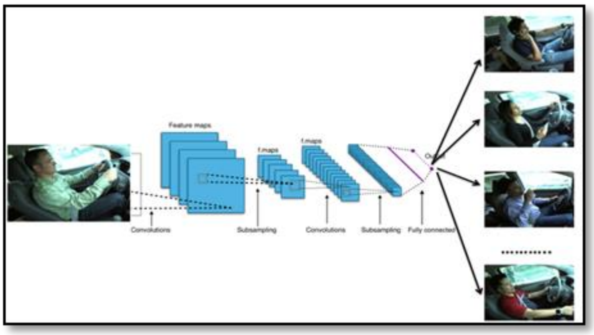


Figure 1. CNN architecture

3.3 ResNet50 architecture

The ResNet50 introduces deep residual learning framework to handle the issue of explosive/vanishing gradients. Shortcut connections are introduced that primarily carry out identity mappings. Fig 2 represents the ResNet50 architecture.

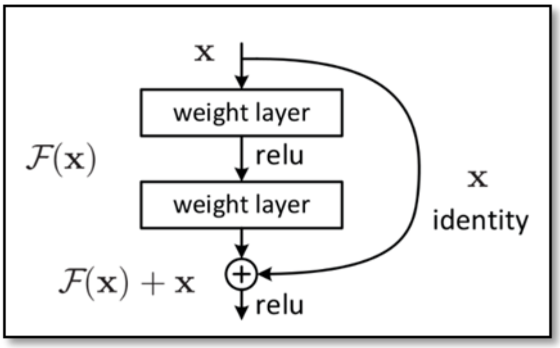

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 2. Resnet50 architecture

Architecture designated the residual mapping as H(x) and allowed the non-linear layers to fit a different mapping, F(x):=H(x)-x, so that the original mapping, as seen in the above figure, became H(x):=F(x)+x. These identity mappings have the advantage of not adding any extra parameters to the model and reducing calculation time.

3.4 Dataset description

For a public Kaggle competition, State Farm provided the data. For the benefit of the research community, Kaggle regularly sponsors machine learning competitions with sizeable cash prizes. The dataset includes of 20,180 training and 5045 testing photos (640 x 480 full colour) of individuals engaged in either attentive driving or one of nine different distracted driving behaviours. An example input images is shown in Fig 3. The training images are supplied with correct labelling and the task is to make the most accurate multi-class classifications possible.

There are ten classes in total:
c0: attentive driving
c1: texting with right hand
c2: talking on the phone with right hand
c3: texting with left hand
c4: talking on the phone with right hand
c5: operating the radio
c6: drinking from the plastic cup
c7: reaching the backseat
c8: setting hair and applying makeup
c9: talking to passengers sitting behind.

The training images were divided into train and validation sets in order to assess the models' correctness. It was crucial to select the photographs of a particular driver from the training set to serve as the validation images since the images are highly associated and each driver only appears in either the training set or the test set. Only then can these validation images be separate from the other drivers' remaining training photos, preventing an artificially high test accuracy. All of the photos from the three randomly selected drivers—roughly 10% of the entire training set—were first split out as the validation set.

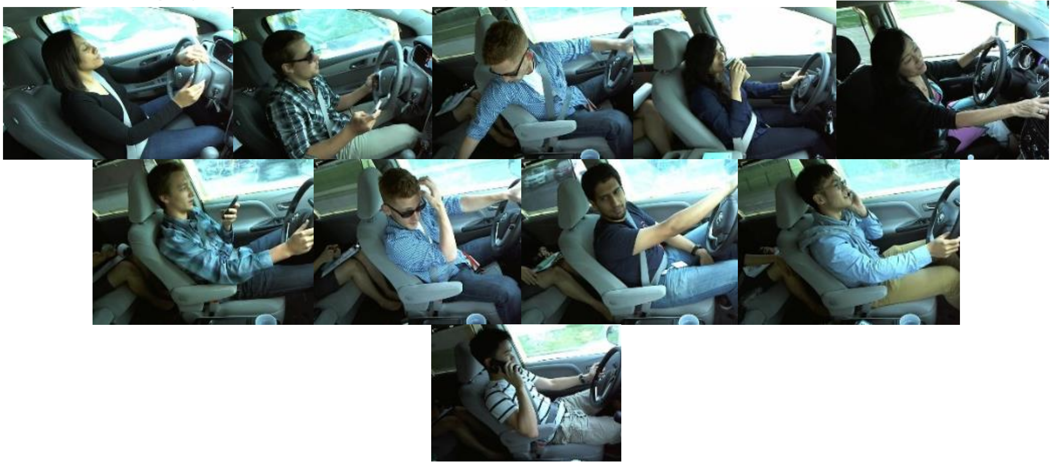


Figure 3. Dataset Description

4. Proposed system

Convolutional Neural Network and its features is used to tackle the problem of classifying the driver’s state. The system proposes a method for detecting driver’s distracted states using deep learning model (RESNET). In system we will be using ResNet-50 architecture, which classifies the given image/video into one of the given classes. Fig 4 represents system design of our model.

4.1 Data preprocessing

Pre-processing is the first step in deep learning workflow to prepare raw data in a format the network can accept. Firstly, the collected dataset will be in different dimensions, so we need to change the dimension of images to a standard form of size 224x244. Initially, the images are resized from 640 x 480 to 224 x 224 pixels. The images are divided into training and test sets.

4.2 Building model and Training

The ResNet model is built and trained on large sets of labeled data. While training, the model tries to learn directly from the data provided. Later, we test the trained model using the test dataset.

Convolution Layer

Convolutional neural networks' central element, the convolutional layer, is always their first layer. It searches the photos supplied as input for a specific set of driver features.

Pooling Layer

This kind of layer is frequently positioned between two convolutional layers. It applies the pooling procedure to each of the several feature maps it gets. The pooling method involves shrinking the size of the State Farm photos while keeping their crucial details.

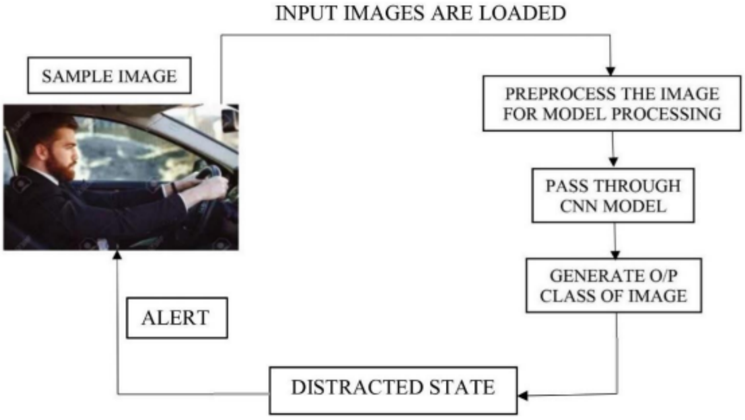


Figure 4. System Design

4.3 Performance and Prediction of Driver state

In this performance module, the performance of trained ResNet model using evaluation criteria such as accuracy score and classification error is shown below. In prediction module, we use trained model to predict whether the uploaded image/video detects the driver state is safe or unsafe. If it’s unsafe, the system creates an alarm signal and respective message will be sent to the organization. System graphical user interface (GUI) is shown in Fig 5, which represesnts both frontend and backend process.

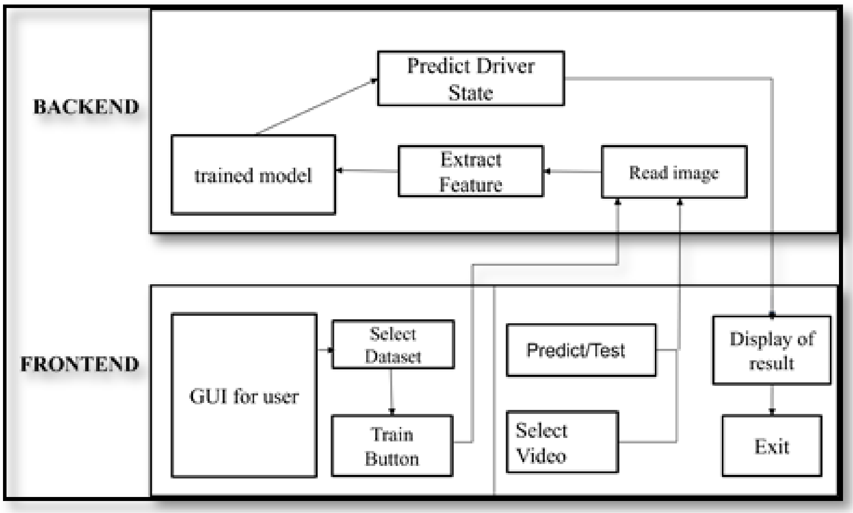


Figure 5. System GUI

5. Results and Discussion

We used the State Farm Distracted Driver Detection dataset supplied by Kaggle to assess the effectiveness of our suggested strategy in identifying distracted driving behaviours. We use the overall accuracy to gauge and assess performance. The percentage of all successfully categorised photos divided by the total number of test samples yields the overall accuracy.

Figure 6 displays the training and validation accuracy curve of our suggested deep learning architecture while using 80% of the training data. When we increase the size of the training data, the accuracy rises. When we use 80% of the data for training, Figure 7 displays the confusion matrix for ten classes of distracted states of our suggested deep learning architecture. And also represents bar graph of our dataset. On model training and testing, the overall accuracy is about 98%, and loss is 1.8%. The validation accuracy is 98.87% and validation loss is 1.13%.

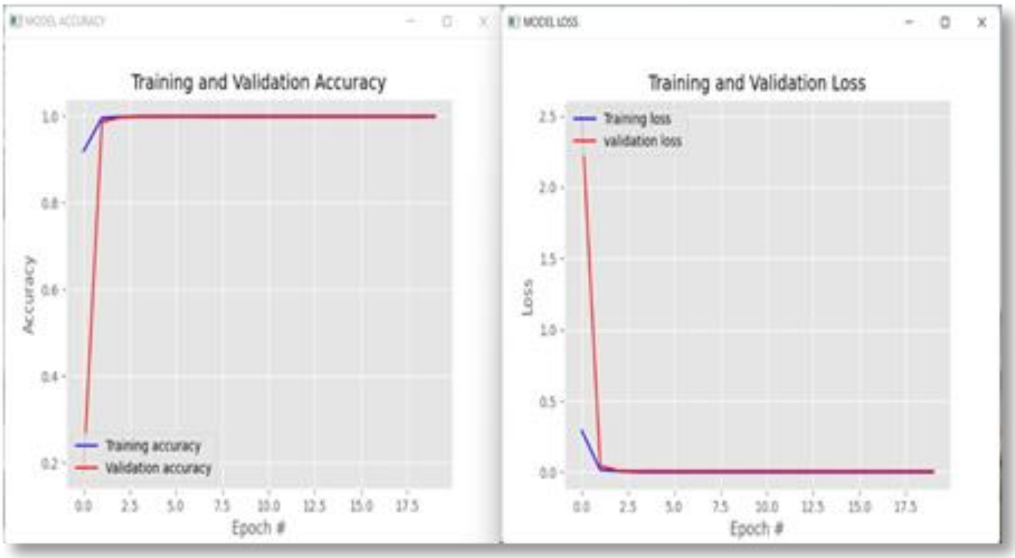
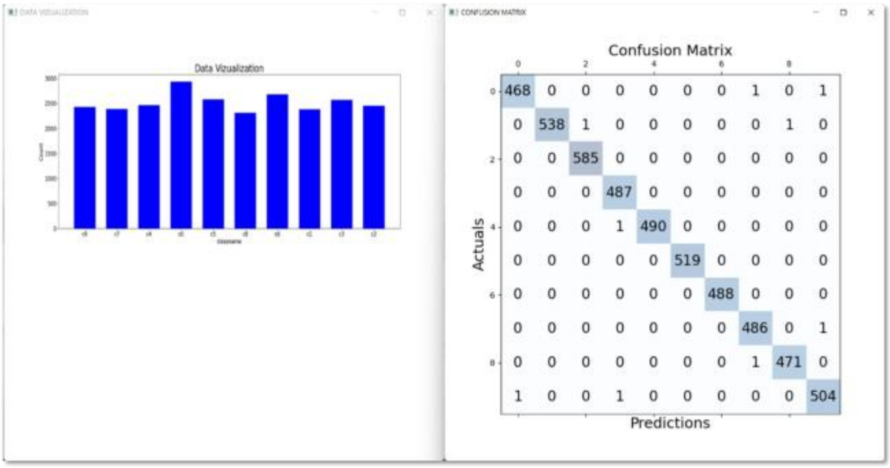


Figure 6. Accuracy and Loss



	0		2		4		6		8	
0	468	0	0	0	0	0	0	1	0	1
	0	538	1	0	0	0	0	0	1	0
2	0	0	585	0	0	0	0	0	0	0
	0	0	0	487	0	0	0	0	0	0
4	0	0	0	1	490	0	0	0	0	0
	0	0	0	0	0	519	0	0	0	0
6	0	0	0	0	0	0	488	0	0	0
	0	0	0	0	0	0	0	486	0	1
8	0	0	0	0	0	0	0	1	471	0
	1	0	0	1	0	0	0	0	0	504

Figure 7. Bar graph and Confusion matrix

6. Conclusion

Live detection of driver’s behaviour is one of the challenging problems faced by automobile companies. In the recent time, as radios, mobile phones and other smart gadgets have become common, accidents occurring due to usage of these devices while driving has increased manifold. We investigate the distracted driver’s posture as a part of human action recognition to recognize the driver’s behaviour. The objective is to classify the images as accurately as possible. We propose a method where the algorithm classifies different states of the driver and if it senses that the driver is exhibiting unsafe posture, it will ring an alarm and send a message to the driver’s manager.

Our approach was verified with State Farm’s dataset on Kaggle’s platform. The images in the dataset were taken from a dashboard mounted camera which captured the various postures of the driver. Finally, our proposed system has given encouraging results for both the state farm dataset as well as our own testing images.

7. References

[1] A. Eriksson and N. A. Stanton, “Takeover time in highly automated vehicles: Noncritical transitions to and from manual control”, Human Factors: The Journal of the Human Factors and Ergonomics Society, vol. 59, no. 4, **(2017),** pp. 689–705**.**

[2] R. Parasuraman and D. H. Manzey, “Complacency and bias in human use of automation: An attentional integration”- Human factors, J. Ergonom. Soc., vol. 52, no. 3, June **(2019),** pp. 381410.

[3] Yang Xing, Chen Lv, Huaji Wang, Dongpu Cao, Efstathios Velenis, Fei-Yue Wang, “Driver Activity Recognition for Intelligent Vehicles: A Deep Learning Approach”, IEEE Transactions on Vehicular Technology, Volume 68, Issue 6, **(2019)**.

[4] Monagih. Alkinani, Wazir Zada khan and Quratulain Arshad, “Detecting Human Driver Inattentive and Aggressive Driving Behavior using Deep Learning”, IEEE Access on Intelligent transport system, vol. 8, **(2020)** , pp. 105008-105030.

[5] Chaoyun Zhang, Rui Li, Woojin Kim, Daesub Yoon and Paul Patras, “Driver Behavior Recognition via Interwoven Deep Convolutional Neural Nets with Multi-Stream Inputs”, 2020 IEEE Access on, vol. 8, **(2020)** , pp. 191138-191151**.**

[6] Binbin Qin, Jiangbo Qian, Yu Xin, Baisong Liu, and Yihong Dong, "Distracted Driver Detection Based on a CNN with Decreasing Filter Size", IEEE Transactions on Intelligent Transportation Systems, Volume 88, Issue 5, **(2021)**.

[7] Duy Tran, Ha Manh Do, Weihua Sheng, He Bai, Girish Chowdhary, "Real-time detection of distracted driving based on deep learning", IEEE Xplore on Intelligent Transport System, Vol. 12, Issue 10, **(2018),** pp. 1210-1219.

[8] Ankit Pala, Subasish Kar and Manisha Bhartia, "Algorithm for Distracted Driver Detection and Alert Using Deep Learning", Springer on Optical Memory and Neural Networks, Vol. 30, No. 3, **(2021),** pp. 257–265.

[9] Z. Chen, J. Yu, Y. Zhu, Y. Chen, and M. Li, "Abnormal driving behaviours detection and identification using smartphone sensors", Proceedings with 12th Annual IEEE International Conference Sensor, Communication and Networks. (SECON), pp. 524532, **(2015)**, June.

[10] R. Chhabra, S. Verma, and C. R. Krishna, "Detecting aggressive driving behaviour using mobile smartphone", Proceedings of 2nd International Conference on Communication, Computing and Networking, Springer, **(2019)**, pp. 513521.