



Article

Advancing Road Safety: A Comprehensive Evaluation of Object Detection Models for Commercial Driver Monitoring Systems

Huma Zia ^{1,*}, Imtiaz ul Hassan ², Muhammad Khurram ², Nicholas Harris ³, Fatima Shah ² and Nimra Imran ²

¹ College of Engineering, Abu Dhabi University, Abu Dhabi 59911, United Arab Emirates

² Smart City Lab, NCAI (National Center of Artificial Intelligence), NED University of Engineering and Technology, Karachi 75270, Sindh, Pakistan; mkhurram@neduet.edu.pk (M.K.); fatimashah@neduet.edu.pk (F.S.)

³ School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK

* Correspondence: huma.zia@adu.ac.ae

Abstract: This paper addresses the critical issue of road safety in the indispensable role of transportation for societal well-being and economic growth. Despite global initiatives like Vision Zero, traffic accidents persist, largely influenced by driver behavior. Advanced driver monitoring systems (ADMSs) utilizing computer vision have emerged to mitigate this issue, but existing systems are often costly and inaccessible, particularly for bus companies. This study introduces a lightweight, deep-learning-based ADMS tailored for real-time driver behavior monitoring, addressing practical barriers to enhance safety measures. A meticulously curated dataset, encompassing diverse demographics and lighting conditions, captures 4966 images depicting five key driver behaviors: eye closure, yawning, smoking, mobile phone usage, and seatbelt compliance. Three object detection models—Faster R-CNN, RetinaNet, and YOLOv5—were evaluated using critical performance metrics. YOLOv5 demonstrated exceptional efficiency, achieving an FPS of 125, a compact model size of 42 MB, and an mAP@IoU 50% of 93.6%. Its performance highlights a favorable trade-off between speed, model size, and prediction accuracy, making it ideal for real-time applications. Faster R-CNN achieved an FPS of 8.56, a model size of 835 MB, and an mAP@IoU 50% of 89.93%, while RetinaNet recorded an FPS of 16.24, a model size of 442 MB, and an mAP@IoU 50% of 87.63%. The practical deployment of the ADMS on a mini CPU demonstrated cost-effectiveness and high performance, enhancing accessibility in real-world settings. By elucidating the strengths and limitations of different object detection models, this research contributes to advancing road safety through affordable, efficient, and reliable technology solutions.



check for
updates

Academic Editor: Laura Eboli

Received: 7 October 2024

Revised: 6 December 2024

Accepted: 17 December 2024

Published: 1 January 2025

Citation: Zia, H.; Hassan, I.u.; Khurram, M.; Harris, N.; Shah, F.; Imran, N. Advancing Road Safety: A Comprehensive Evaluation of Object Detection Models for Commercial Driver Monitoring Systems. *Future Transp.* **2025**, *5*, 2. <https://doi.org/10.3390/futuretransp5010002>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: advanced driver monitoring systems; driver behavior detection; deep learning; object detection models; real-time monitoring; transportation safety

1. Introduction

Transportation plays a pivotal role in enhancing individual and societal well-being, fostering economic growth, and elevating overall quality of life [1]. Nonetheless, it is not without its grave implications, particularly the distressing prevalence of traffic accidents that carry the potential for severe harm or loss of life. In response to these pressing concerns, governments and policymakers have embarked on initiatives aimed at enhancing road safety.

One noteworthy initiative is “Vision Zero”, which originated within the Swedish parliament and has garnered international recognition [2]. This paradigm strives to create a world where no traffic-related fatalities or serious injuries are deemed acceptable. Vision

Zero has transcended borders and found adoption in diverse nations, including developing nations like Pakistan, as well as developed nations like the UAE [3]. Its implementation seeks to instill comprehensive road safety practices and curtail the occurrence of fatal accidents on a global scale. The presented work is inherently aligned with the objectives of Vision Zero, which seeks to eliminate traffic-related fatalities and severe injuries.

While driving is a routine activity, it carries inherent risks for both drivers and others on the road. These risks stem from factors like road layout, conditions, surroundings, and driver behavior [4]. Notably, driver behavior remains an unpredictable aspect influenced by the interplay of the driver, environment, and vehicle. Research has indicated that accidents tend to decrease when co-passengers alert drivers to unseen dangers, such as instances of inattentiveness that may lead to accidents, with rates ranging from 30% to 43% [5].

The rapid evolution of computer vision technologies has led to the development of automatic driver monitoring systems, adept at detecting instances of driver distraction. These sophisticated monitoring systems can seamlessly integrate with automated control mechanisms, effectively notifying drivers about their distracted behaviors [6].

Leading high-tech car manufacturers have started integrating advanced driver monitoring systems into their vehicles. In 2018, Volvo introduced an innovative Driver Alert Control (DAC) system, utilizing a camera to detect the road's side markings [7]. This camera then compares the detected road section with the driver's steering wheel movements, effectively monitoring their interaction with the road.

Similarly, Mercedes-Benz has implemented an advanced Attention Assist system [8]. This system relies on a specialized steering wheel sensor, meticulously recording the driver's steering wheel movements and speed. Through sophisticated analysis of data from steering wheel sensors, the Attention Assist algorithm gains insights into the driver's personal driving habits, detecting potential signs of drowsiness and promptly notifying drivers, thereby preventing momentary lapses in attentiveness.

In the event of any deviation or change in the driver's established driving style, the system promptly issues alerts [8]. By seamlessly integrating technology and behavior analysis, Mercedes-Benz's Attention Assist contributes significantly to road safety by detecting potential signs of drowsiness and promptly notifying drivers, thereby preventing momentary lapses in attentiveness.

Both Volvo and Mercedes-Benz have demonstrated remarkable strides in enhancing road safety through the development of advanced driver monitoring systems. These systems leverage cutting-edge technology to continuously assess driver behavior, promptly alerting drivers to potential risks and ensuring a heightened level of vigilance on the road.

However, a major issue with most of the systems developed by the automotive industry is their high cost [9]. Due to their dependence on expensive sensors, these smart tracking systems are only available in high-end vehicles, making them inaccessible to the general public. This is perhaps why, in recent years, multiple academic studies have attempted to develop accurate yet cheap drowsiness detection systems for real-world use.

Previous work [10,11] in the field of driver attentive state detection has predominantly focused on utilizing a single feature to assess driver behavior and attention level. While these studies have provided valuable insights, there are limitations to relying solely on a single feature. Using only one feature restricts the analysis to a narrow aspect of driver behavior, providing an incomplete understanding of the driver's attentive state. Single-feature-based models tend to perform well in controlled environments or specific scenarios, but their effectiveness may diminish when applied to diverse real-world situations.

The paper presents a deep-learning-based lightweight advanced driver monitoring system tailored to address the practical challenges of real-time driver behavior monitoring. This system is designed to detect various aspects of driver behavior on the road, facilitating

enhanced safety measures and assisting autonomous driving functionalities. The main contributions of this work include:

- A lightweight deep learning model is developed for a commercial advanced driver monitoring system, collecting real-time observable driver information to assess their ability in dynamic driving tasks.
- The research utilizes a meticulously curated dataset, representing diverse demographics, to enhance model accuracy across different population groups. Varied lighting conditions are considered during data collection, strengthening the model's robustness and adaptability in diverse driving environments.
- Three state-of-the-art object detection models (Faster R-CNN, RetinaNet, and YOLOv5) are employed to detect various drivers' behavioral features. The YOLO model's superior performance, particularly its inference time for frame extraction and detection, as well as its model size, makes it an ideal choice for the driver behavior monitoring system.
- The model is implemented on a mini CPU to create a cost-effective, commercially viable, high-performance, and low-power standalone driver behavior monitoring system.

2. Literature Review

The literature review begins by providing an overview of the techniques utilized in driver behavior detection systems, starting with the use of inertial and biological sensors. It then delves into advancements in computer vision techniques, with a particular emphasis on the integration of deep learning models. Within this context, the literature highlights three specific deep learning models, which are used in the current study. These models are described in terms of their accuracies and distinctive features, showcasing their potential for improving the performance of driver behavior detection systems.

2.1. Inertial and Biological Sensors

A system [12] was designed to identify a driver's inattention due to secondary task distraction. Three states were identified: handling a CD, reaching for an object on the back seat, and checking the speedometer and mirrors. Data for the system were collected using head-mounted inertial sensors, including an accelerometer, gyroscope, and magnetometer, and drivers were asked to drive on a specific path. To classify the captured data, three machine learning classifiers were trained: support vector machine (SVM) with linear and radial basis function (RBF) kernels, k-nearest neighbor (k-NN), and random forest (RF). With the exception of SVM linear, all the employed techniques yielded an accuracy, precision, and recall exceeding 96%. However, the controlled setup of the study makes it impractical to apply findings to real-world driving conditions.

EEG and EOG recordings [13] were utilized to estimate the level of vigilance inside the vehicle. Long short-term memory (LSTM) combined with capsule feature extraction was employed to learn representative features. The system's performance was evaluated using two metrics: root mean square error (RMSE) and Pearson correlation coefficient (PCC). The calculated values for RMSE and PCC were 0.0295 and 0.9887, respectively. Despite the model's high accuracy in detecting the driver's state, practical applications do not commonly employ physiological measures such as EEG and EOG due to their intrusive nature. Additionally, the implementation of a detection system that requires hardware for direct contact with the driver's body may cause discomfort during prolonged periods of use.

Another approach [14] was implemented to develop a real-time driver distraction detection system by collecting eye and head movement measurements. Data from the eye and head movement tracker were used to evaluate the performance of Laplacian support

vector machines (LapSVM) and a semi-supervised extreme learning machine (SS-ELM) model. SS-ELM achieved the greatest accuracy of 97.2% and G-mean of 0.959. However, it is important to note that the accuracy of driver distraction detection decreases when there are changes in eye movement behavior, such as when a different driver is involved. This indicates that the system's performance may be influenced by individual variations in eye movement patterns.

Accelerometer and gyroscope sensors [15], embedded in modern smartphones, were utilized to detect distracted driving behavior. This behavior includes activities such as making phone calls, sending text messages, and reading while driving. A random-forests-based algorithm was employed to classify distracted driving activities in real time. The system was extensively evaluated across multiple metrics, environments, and testing strategies, yielding favorable results. Despite the outstanding classification results (85% precision and 84% recall measures for both sensors combined), the entire study was conducted in a controlled environment. This limited scope neglects various possible scenarios, rendering it non-feasible for real-world conditions. Additionally, smartphone-based detection methods may encounter limitations when drivers forget to carry their smartphones, when the smartphones are switched off, or when the cause of distraction is unrelated to smartphone usage.

When drivers utilize electronic devices or become distracted by the surrounding scenery, they divert their attention away from the road. Research work based on gaze monitoring sensors has been conducted to predict drowsiness, fatigue, and distraction. An SMI ETGTM eye tracker [11] was employed to obtain drivers' gaze behavior parameters in a driving simulation platform. The drivers were assigned certain secondary tasks to assess differences in visual characteristics during task performance. In another study, gaze information [10] was analyzed using a front camera to extract features such as facial landmarks, head pose, and iris centers. LSTM was utilized for driver behavior monitoring and prediction. However, while gaze monitoring provides sufficient information about driver visual behavior, it is insufficient to accurately determine driver distracted states. In addition to monitoring visual behavior, analyzing a driver's posture can provide valuable insights for detecting distracted states. The Kanade–Lucas–Tomasi (KLT) point tracker [16] was used to track body parts, including the hand, lips, and forehead, while driving. The proposed feature sets were combined with the kernel SVM technique to detect and classify different types of distractions during driving. Different prototypes have been developed, but none have been commercialized for the automotive industry due to the variance of different types of sensors, which makes them unpredictable and unsuitable for real-world conditions.

2.2. Deep Learning Models

Using footage from video cameras provides a more flexible and effective approach to monitoring driver behavior. By incorporating computer vision (CV) methods, video camera footage can offer real-time and highly accurate information about the driver. Object detection models such as Faster R-CNN, CNN, YOLO, SSD, or RetinaNet can be trained to identify specific objects of interest, including pedestrians, mobile devices, seat belts, cigarettes, vehicles, or weapons, in real-time video footage. Additionally, posture detection and face identification models can be trained to detect signs of driver fatigue and inattentiveness.

2.3. Faster RCNN

Deep ConvNets [17] are widely used for object detection and image classification due to their higher accuracy compared to previous models such as VGGNets, ResNets, Inception networks, and DenseNet. One notable architecture is R-CNN, which utilizes a

deep ConvNet to recognize object proposals (potential regions of interest). Although it achieves high accuracy, it suffers from time and space inefficiencies. The system takes a long time and requires a large amount of storage space as it extracts features from each image and saves them to a hard disk. The detection process alone takes 47 s for a single image. Fast R-CNN [18] significantly improves the detection speed to 0.3 s per image by incorporating an ROI-pooling layer. However, the region proposal step, which is not part of the network architecture, becomes a bottleneck in the system. Consequently, the overall solution becomes suboptimal, and the network relies on external methods for region proposal.

The drawback of Fast R-CNN is addressed by Faster R-CNN [19], which introduces the region proposal network (RPN). The RPN is implemented as a fully convolutional network that predicts object boundaries and objectness scores. It achieves translation invariance using anchors with different scales and ratios. By integrating the deep VGG-16 model, the entire system can efficiently perform the proposal and detection process in just 0.2 s [20]. One study [21] proposes an ensemble learning approach based on deep learning techniques to detect distracted drivers. It detects 10 distracted states, including talking on the cell phone, texting, tapping on the cell phone screen, smoking, drinking, eating, taking hands off the steering wheel, mending facial hairs, talking to passengers, and taking eyes off the road. By fine-tuning the Faster R-CNN model and extracting pose points from the driver's posture, the approach achieves high accuracy (97.7% validation accuracy). The model focuses on objects directly associated with distraction and calculates interactive associations using the intersection over union metric. It achieves an accuracy of 92.2%, surpassing RCNN and Fast R-CNN. To ensure its practicality, the study should evaluate the model's real-time performance, considering computational efficiency and response time. Another study mentions an improved Faster R-CNN model [22] designed specifically for small object detection. The approach introduces novel techniques for bounding box regression and RoI pooling to address positioning deviation issues. To ensure robustness, the study curated the TT100K dataset (Tsinghua-Tencent 100K) which includes a diverse range of traffic signs, accounting for variations in luminance and weather conditions. The model incorporates multi-scale convolution feature fusion and an improved NMS (non-maximum suppression) algorithm for accurate recognition. Results demonstrate high performance on traffic signs, achieving a recall rate of 90% and an accuracy rate of 87%. This indicates the effectiveness of Faster R-CNN for small object detection. However, further research is necessary to evaluate its performance on different objects and domains, taking into account computational efficiency and potential limitations.

2.4. YOLO

Object detection is a challenging problem in computer vision, and deep learning has significantly improved the performance of object detectors in tasks such as classification, localization, and segmentation. Two-stage detectors, such as Faster R-CNN, utilize complex architectures for selective region proposals, while single-stage detectors like YOLO [23] employ simpler architectures to process all spatial regions in one shot. While two-stage detectors generally achieve higher detection accuracy, single-stage detectors, particularly YOLO-based models, offer faster inference times [24]. The trade-off between accuracy and speed has made YOLO popular in various applications. For example, although YOLO may have a detection accuracy of 63.4 compared to Fast R-CNN's detection accuracy of 70, it can perform inference around 300 times faster.

One study [25] presents solutions for object detection and tracking in an autonomous driving scenario. The study compares the performance of state-of-the-art object detectors, namely YOLOv5, Scaled-YOLOv4, and YOLOR, trained on the BDD100K dataset, which

includes out-cabin objects such as cars, pedestrians, lanes, and traffic lights. The algorithm is deployed on the NVIDIA Jetson AGX Xavier Edge Device. Real-time inference capabilities are evaluated using DeepStream technology, and different object trackers (NvDCF and DeepSORT) are compared using the KITTI tracking dataset. The proposed solution achieves a detection interval of one with a frame rate of 33.3 FPS and a detection interval of one with a frame rate of 17 FPS using the YOLOR-CSP architecture with a DeepSORT tracker. Another study [26] investigates the application of YOLO-based deep learning models for in-cabin monitoring and occupant detection in driving scenarios. The study utilizes a fisheye-lens camera and RGB-format images as inputs and evaluates various YOLO models, including YOLOv3-tiny, YOLOv3-tiny-3l, YOLO-fastest, YOLO-fastest-xl, and YOLO-fastest-three scales. The results demonstrate that the YOLO-fastest-three scales model achieves the highest F1-score (95.89%) and mAP (97.16%), while the YOLO-fastest-xl model exhibits the lowest false negative rate (2.63%). The proposed design executes at up to 30 FPS on a GPU-based embedded device.

2.5. RetinaNet

RetinaNet, a pivotal model in object detection, has captured substantial attention in the computer vision community. Lin et al. [27] introduced RetinaNet in 2018, proposing a fusion of feature pyramid networks (FPN) and a novel loss function called focal loss. Focal loss addresses class imbalance, assigning higher weights to hard-to-find objects, particularly small objects, enhancing RetinaNet's accuracy. The combination of FPN and focal loss enables RetinaNet to achieve state-of-the-art performance in both one-stage and two-stage object detection models. It stands out for its proficiency in handling objects of various scale.

In another [28] study, the authors address the challenge of accurately obtaining the number of wheat ears, a crucial indicator for wheat production and yield estimation. The study focuses on comparing the performance of faster regions with convolutional neural networks (Faster R-CNN) and RetinaNet in predicting the number of wheat ears at different growth stages and under diverse conditions. The results, utilizing the Global WHEAT dataset for recognition, reveal that the RetinaNet method and the Faster R-CNN method achieve average accuracies of 0.82 and 0.72, respectively, with RetinaNet demonstrating higher recognition accuracy. Moreover, when utilizing collected image data for recognition, the R2 values after transfer learning for RetinaNet and Faster R-CNN are 0.9722 and 0.8702, respectively, indicating superior recognition accuracy for the RetinaNet method across different datasets.

Another study [29] focuses on the development and evaluation of an object detection system for detecting storm-drains and manholes in the streets of Campo Grande city, Brazil. Terrestrial images were acquired, and a dataset containing 297 images was created, manually annotated, and divided into training, validation, and testing sets. The RetinaNet object detection method was adopted for its ability to handle class imbalance, with ResNet-50 and ResNet-101 as backbones. The training and validation loss curves indicate successful convergence without overfitting. RetinaNet outperformed Faster R-CNN in terms of average precision (AP), with better results for both manhole and storm-drain classes.

This paper aims to develop a robust model for an advanced driver monitoring system (ADMS) using state-of-the-art models, namely YOLOv5, Faster RCNN, and ResNet, which will be trained and tested on customized datasets. These models will then be compared based on inference time, accuracy, and model size. The findings from this research will contribute to the development of an ADMS capable of effectively detecting and addressing driver behavior and associated potential risks in real time.

3. Materials and Methods

In this section, we delve into the detailed approach employed in our study, which commences with the creation of a diversified dataset encompassing essential features crucial in driver behavior and distraction monitoring, including eye closure, yawning, seat belt detection, smoking, and mobile phone usage. Following this, the preprocessing and annotation of the collected data are conducted using LabelImg. Subsequently, three state-of-the-art machine learning algorithms—Faster R-CNN [19], RetinaNet [27], and YOLOv5 [23]—are employed to comprehensively evaluate object detection models. Our investigation meticulously examines critical performance metrics, including frames per second (FPS), model size, and learning rate, to offer valuable insights into the trade-offs associated with each model. Figure 1 illustrates the process flow diagram adopted for the development of the ADMS.

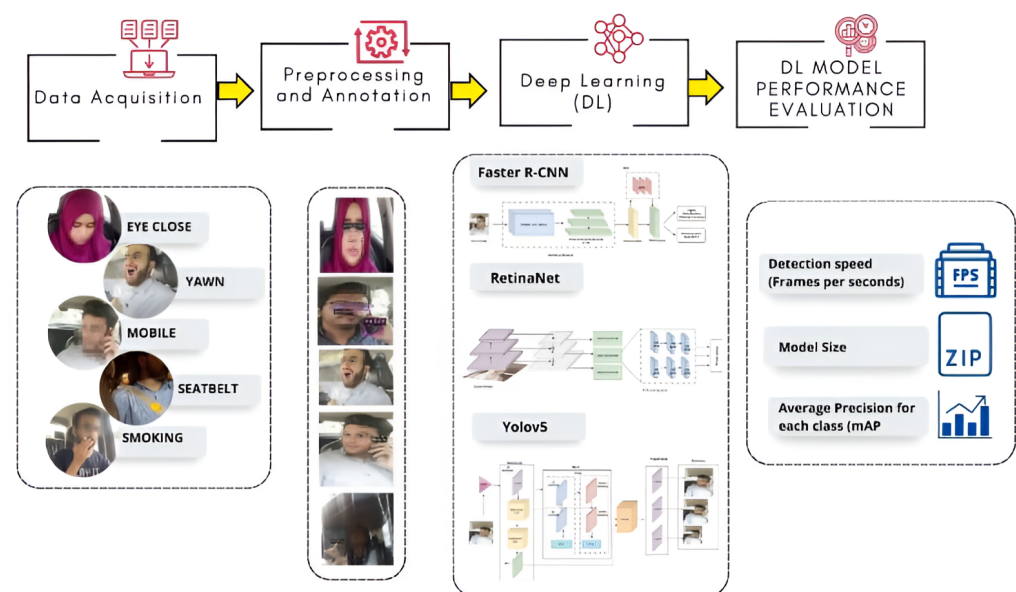


Figure 1. Visual representation outlining the development process of the ADMS (faces in the images have been blurred to protect the privacy of individuals in accordance with ethical publishing standards).

3.1. Dataset Description

Driver behavior plays a critical role in road safety, with dangerous behaviors often leading to fatal accidents. Figure 2 provides insights into roadside interactions of drivers, focusing on five key behavioral features.

The dataset used for this research was collected via a dashboard camera installed in a car, capturing the in-cabin behaviors of drivers. Careful consideration was given to encompass diverse aspects such as religious diversity, ethnicity, and skin color. This inclusive approach aimed to ensure a comprehensive representation of various demographic groups, thereby enhancing the model's accuracy and applicability across different populations. Additionally, varied lighting conditions were considered during data collection. The objective was to ensure that the dataset captured variations in lighting scenarios commonly encountered during driving. This consideration aimed to enhance the robustness and adaptability to diverse driving environments. Figure 3 illustrates a diversified dataset, encompassing each feature while representing varied lighting conditions and demographic populations.

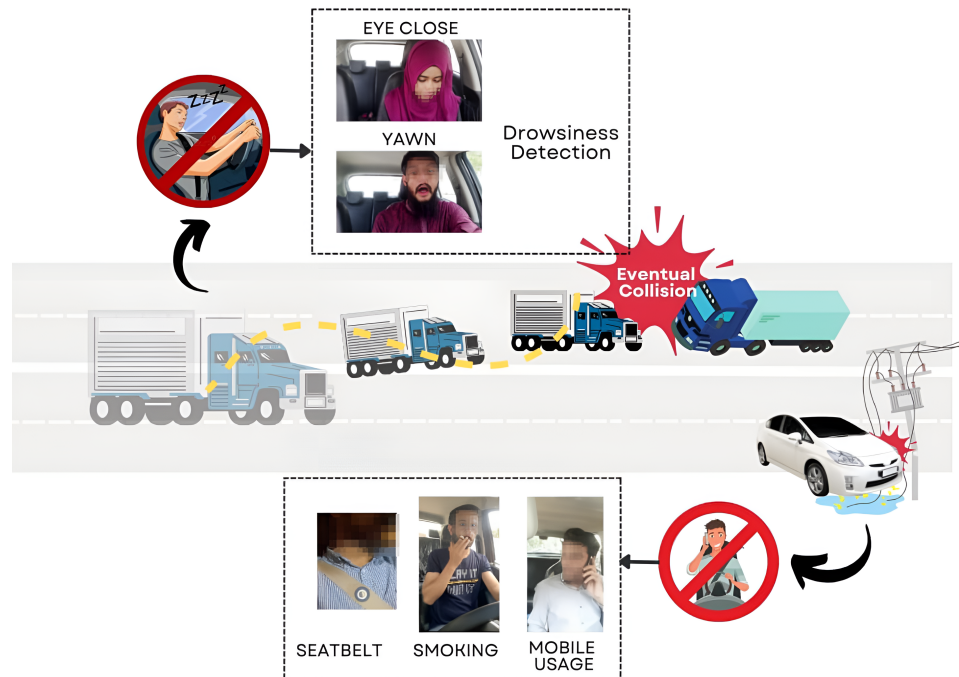


Figure 2. Illustration depicting roadside interactions and highlighting five crucial behavioral features for driver monitoring (faces in the images have been blurred to protect the privacy of individuals in accordance with ethical publishing standards).



Figure 3. Visuals of a diversified dataset showcasing each feature, reflecting varied lighting conditions and demographic populations (faces in the images have been blurred to protect the privacy of individuals in accordance with ethical publishing standards).

The dataset encompasses 1050 images for capturing instances of eye closure, 550 images for yawning, 760 images for seat belt detection, 570 images portraying instances of smoking, and 1900 images for mobile phone usage detection. The distribution of classes in the dataset is illustrated in Figure 4. The participants in this study were research assistants from the SmartCity Lab at the National Center for Artificial Intelligence (NCAI), NED University of Engineering and Technology, Karachi. All participants provided in-

formed consent prior to their involvement in the data collection process. For seat belt detection, a unique challenge was encountered as seat belts often became camouflaged with the clothing worn by drivers. To address this issue, a distinctive sticker logo was designed and affixed to the seatbelt, facilitating improved visibility for the model during training and inference. Given that the system is designed for commercialization, the entire system, including the designed sticker for seatbelt enhancement, will be installed. This diverse and comprehensive collection of images for each behavior category serves as a robust foundation for training and enhancing the accuracy and efficacy of the ADMS in detecting and interpreting crucial driver behaviors. Additionally, the dataset comprises a total of 3544 training images, 532 testing images, and 890 validation images.

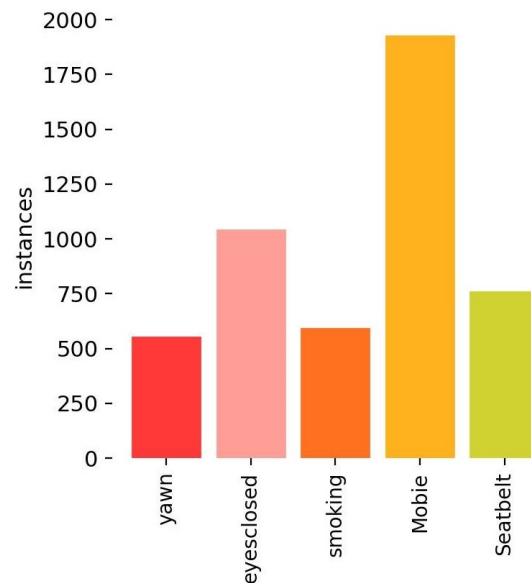


Figure 4. Bar graph illustrating the distribution of classes in the dataset, showcasing instances of eye closure, yawning, seat belt detection, smoking, and mobile phone usage.

3.2. Experimental Setup and Training

The experimental setup incorporates a robust computational infrastructure featuring the NVIDIA RTX A4000 GPU with 16GB memory and 32GB RAM. This selection is strategic, aiming to optimize the training and inference efficiency of Faster R-CNN, RetinaNet, and YOLOv5, particularly in the context of intricate tasks such as real-time driver monitoring.

The training configurations for the specified models are as follows. The Faster R-CNN model with X-101 32x8d FPN 3x was trained with a learning rate of 0.0001 for three epochs. Similarly, the RetinaNet_R_101_FPN_3x model used a learning rate of 0.0001 and was trained for three epochs. On the other hand, the YOLOv5 Medium model employed a learning rate of 0.0001 and underwent a more extensive training process, spanning 19 epochs. The architecture of each model is described in the subsequent sections.

3.2.1. Faster R-CNN

The Faster R-CNN model [19] with the X-101 32x8d FPN backbone, is a sophisticated architecture for precise object detection in images. At its core, the model utilizes an extended variant of the ResNet architecture, known as X-101, which features a deep structure with 32 layers and an 8× width expansion at each layer. This backbone network acts as a feature extractor, capturing intricate details from input images. The integration of a feature pyramid network (FPN) enhances the model's ability to detect objects at multiple scales, facilitating the recognition of objects of varying sizes within an image.

Complementing the backbone is the region proposal network (RPN), which efficiently generates region proposals—potential bounding boxes containing objects [19]. Operating on the feature maps provided by the FPN, the RPN identifies candidate regions for further analysis. Subsequently, the regions of interest (ROIs) are selected based on these proposals. The ROI pooling layer ensures consistent feature alignment within these regions, facilitating subsequent layers in performing accurate classification and bounding box regression.

The final stages of the architecture involve refining the proposed regions through bounding box regression and determining the object classes via object classification. Figure 5 illustrates the architecture of the Faster R-CNN model and showcases the intricate network design and components involved in its operation. This comprehensive approach enables the model to not only identify the presence of objects but also precisely localize them within the image.

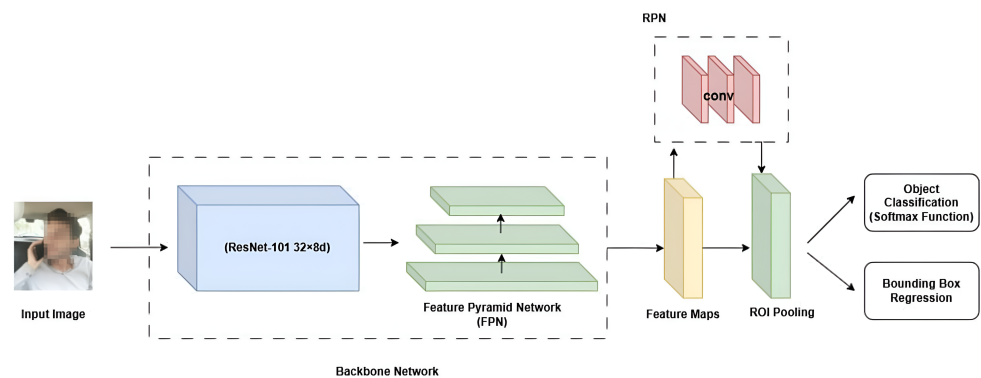


Figure 5. Architecture diagram of the Faster R-CNN model depicting the intricate design and its components (faces in the images have been blurred to protect the privacy of individuals in accordance with ethical publishing standards).

The Faster R-CNN with X-101 32x8d FPN 3x model was chosen for its superior object detection precision, with the depth of the X-101 32x8d FPN backbone and adaptability to various scales through FPN. Its proven track record [ref] in accurately localizing and classifying objects, pre-trained weights, and strong community support make it the optimal choice for achieving high-level accuracy in the advanced driver monitoring system.

3.2.2. RetinaNet

RetinaNet_R_101_FPN_3x [27] represents a configuration of the RetinaNet object detection model. RetinaNet is known for its effectiveness in addressing class imbalance during object detection tasks, thanks to the introduction of focal loss. In this specific configuration, the model employs ResNet-101 (R_101) as its backbone architecture, using the depth and representational power of ResNet networks. A feature pyramid network (FPN) is integrated to create a feature pyramid, enhancing the model's ability to detect objects at various scales. Finally, the model is trained for three epochs, allowing it to learn from the dataset over an extended training period. This configuration is tailored to achieve accurate and robust object detection by combining advanced architectural elements and extended training duration. Figure 6 provides an insightful depiction of the RetinaNet model's architecture, revealing the intricate design and key components essential for object detection tasks.

The RetinaNet_R_101_FPN_3x model was selected for its object detection prowess, specifically its use of the ResNet-101 backbone and feature pyramid network (FPN) for multi-scale feature extraction [27]. This configuration ensures accurate detection across diverse scales, crucial for driver monitoring.

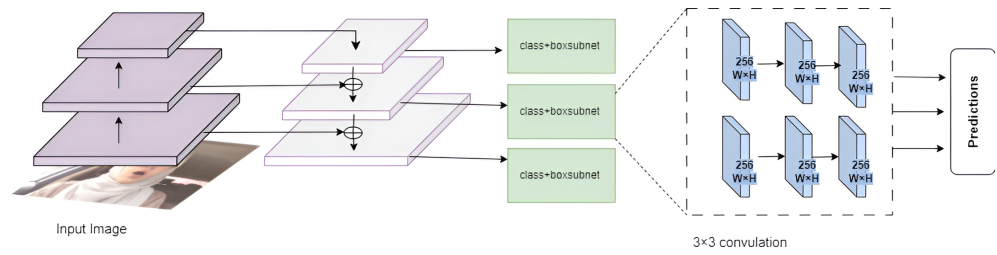


Figure 6. Architecture diagram illustrating the structure and components of the RetinaNet model.

3.2.3. YOLOv5m

The YOLOv5m (medium) [23] architecture is characterized by its utilization of the CSPDarknet53 backbone, an enhanced version of Darknet, for effective feature extraction. Complemented by a feature pyramid network (FPN) and a PANet neck architecture, YOLOv5m excels in multi-scale feature extraction and path aggregation, enhancing its ability to detect objects of varying sizes. Figure 7 gives an in-depth view of the YOLOv5 model’s architecture, presenting its streamlined design and key elements crucial for efficient object detection. The YOLO head is responsible for generating predictions, dividing the input image into a grid, and predicting bounding boxes and class probabilities. YOLOv5m employs anchor boxes and a combination of binary cross entropy (BCE) loss with focal loss during training to address class imbalance and emphasize challenging objects. As a medium-sized variant, YOLOv5m strikes a balance between model size and computational efficiency, making it versatile for applications where real-time or near-real-time inference is essential, such as video analysis, surveillance, and robotics. The architecture’s adaptability, anchored in a robust backbone and advanced feature extraction mechanisms, underscores its effectiveness across diverse object detection tasks.

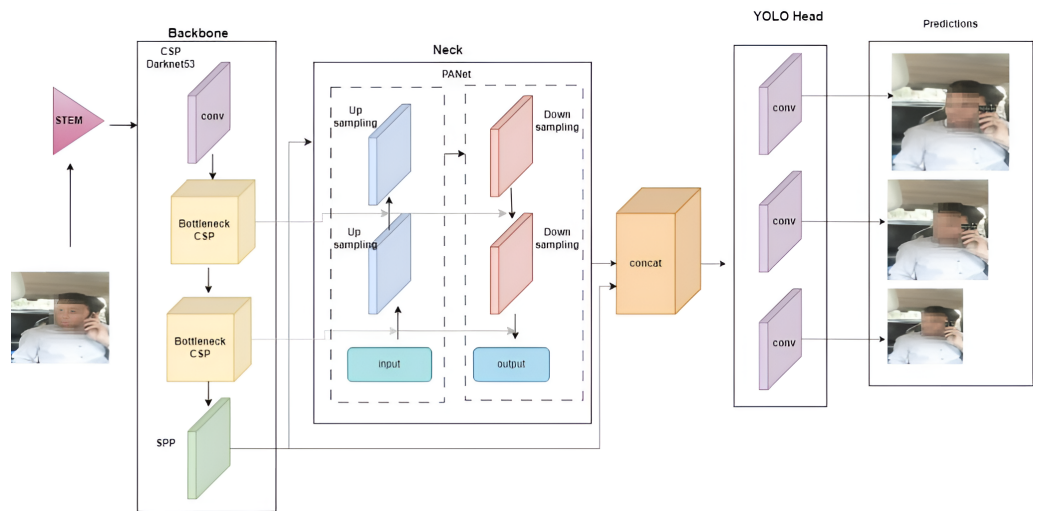


Figure 7. Architecture diagram showcasing the streamlined design of the YOLOv5 model (faces in the images have been blurred to protect the privacy of individuals in accordance with ethical publishing standards).

YOLOv5m is geared towards achieving real-time or near-real-time inference speeds [23], making it well suited for applications such as video analysis, surveillance, and robotics, where low latency is essential.

4. Results

In this section, we present a comprehensive evaluation of three state-of-the-art object detection models, namely Faster R-CNN, RetinaNet, and YOLOv5, for driver monitoring systems. Our investigation focuses on critical performance metrics, including frames per

second (FPS), model size, and learning rate, to provide valuable insights into the trade-offs associated with each model.

4.1. Faster R-CNN

On the test dataset, the Faster RCNN model demonstrated its capability to detect and classify objects with varying sizes and complexities. The model performed exceptionally well in localizing specific behaviors, such as seatbelt usage. The Faster R-CNN model operates at a frames per second (FPS) rate of 8.56. With a model size of 835 MB, it is a relatively large model. The learning rate during training is set at 0.0001, signifying the step size used to adjust model weights during the optimization process. Table 1 presents average precision values against each feature, highlighting the model's performance across different behavioral categories. The table reveals that the model achieved high precision for seatbelt detection, with an average precision value of 70.675. However, performance varied across other features, with eye closed detection achieving an average precision of 51.815, smoking detection at 30.612, mobile phone usage detection at 37.502, and yawning detection at 54.628.

Table 1. Average precision values achieved by the Faster R-CNN model for different behavioral features.

Category	AP
warnings	nan
seatbelt	70.675
eye closed	51.815
smoking	30.612
mobile	37.502
yawn	54.628

4.2. RetinaNet

The per-category bounding box average precision (AP) evaluation reveals noteworthy findings in the model's performance. The model demonstrates higher precision in detecting seatbelt usage (AP: 78.849), emphasizing its effectiveness in accurately identifying instances of this critical safety behavior. However, there is room for improvement in detecting eye closure (AP: 43.786), smoking-related instances (AP: 38.012), and mobile phone usage (AP: 37.655), where precision is moderate. On a positive note, the model excels in detecting yawning behaviors, with a high AP of 60.848. The model's performance could be further optimized for eye closure, smoking, and mobile phone usage detection, providing targeted areas for refinement. Table 2 showcases the average precision achieved by the RetinaNet model across different behavioral categories, further enriching our understanding of its performance.

Table 2. Illustrates the average precision attained by the RetinaNet model across various behavioral categories, offering insights into its performance.

Category	AP
warnings	nan
seatbelt	78.849
eye closed	43.786
smoking	38.012
mobile	37.655
yawn	60.848

4.3. YOLOv5m

The YOLOv5 Medium model showcases exceptional performance in object detection on the test dataset, with an overall precision of 0.876 and a recall of 0.925. Operating at a frames per second rate of 125, the model demonstrates efficiency in real-time processing. Notably, it excels in detecting specific driver behaviors such as yawning, eye closure, smoking, mobile phone usage, and seatbelt usage, achieving high precision and recall scores across these categories. Table 3 showcases the YOLOv5 Medium model’s object detection results for various features, highlighting precision, recall, and mAP@IoU 0.5 scores alongside the number of images and instances. The compact model size of 42 MB enhances its deployability, making it a well-suited choice for applications like driver monitoring systems, where accurate and real-time object detection is paramount. Figure 8 depicts the dynamic progression of precision at a 50% IoU, which stands at 0.876 after each epoch for all behavior classes.

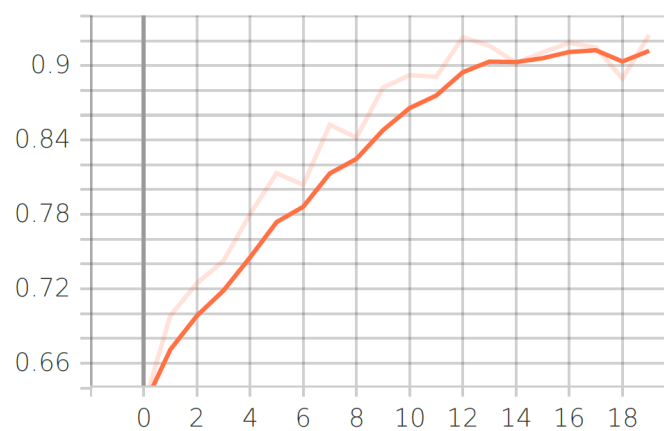


Figure 8. The changing precision over epochs at a 50% IoU threshold, with a stable value of 0.876 for all behavior classes.

Figure 9 illustrates the evolving trend of mAP@ at 50% IoU, achieving a consistent value of 0.936 after each epoch across all behavior classes.

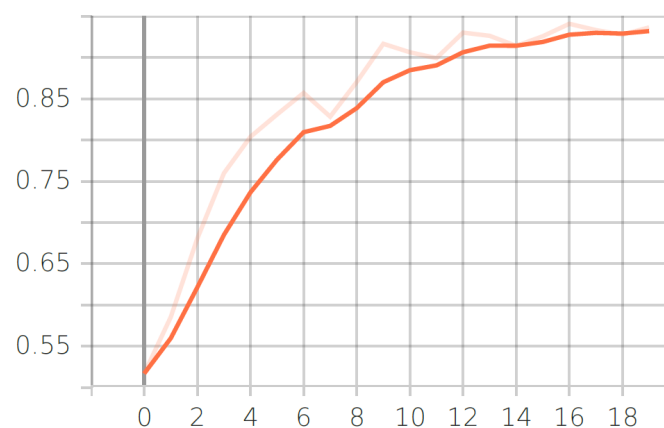


Figure 9. mAP variation at a 50% IoU, remaining steady at 0.936 for all behavior classes after each epoch.

Figure 10 illustrates the changing trend of recall, depicting how it evolves over each epoch at a 50% IoU. Despite this evolution, the recall consistently reaches a value of 0.9256 after each epoch across all behavior classes, highlighting the model’s stability in accurately capturing instances of various behaviors.

These details highlight the trade-offs between model size, processing speed (FPS), and learning rate. YOLOv5m stands out with a substantially higher FPS and smaller model size,

suggesting it might be more suitable for real-time applications where speed and efficiency are critical. However, the final choice depends on the specific requirements and constraints of the driver monitoring system. Figure 11 demonstrates the results of applying YOLOv5 for real-time driver behavior detection using a diverse dataset.

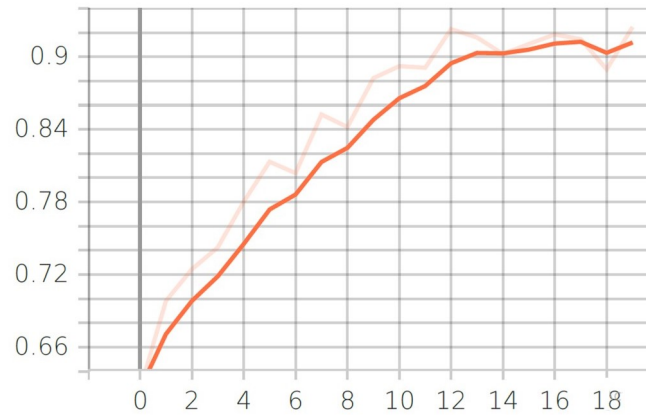


Figure 10. Recall variation over epochs at 50% IoU threshold.

Table 3. YOLOv5m object detection results for different classes, including the number of images and instances, as well as precision, recall, and mAP@IoU 0.5 scores.

Class	Images	Instances	Precision	Recall	mAP@IoU0.5
all	582	573	0.876	0.925	0.936
yawn	582	90	0.97	0.978	0.978
eye closed	582	214	0.857	0.958	0.933
smoking	582	97	0.956	0.918	0.935
mobile	582	74	0.754	0.77	0.843
seat belt	582	98	0.815	1	0.983



Figure 11. Detected driver behavior during driving (faces in the images have been blurred to protect the privacy of individuals in accordance with ethical publishing standards).

In the evaluation of three distinct object detection models for driver monitoring systems, Faster R-CNN, RetinaNet, and YOLOv5, critical performance metrics were scrutinized. Faster R-CNN exhibited a moderate frames per second (FPS) rate of 8.56, accompanied by a larger model size of 835 MB, utilizing a common learning rate of 0.0001. Conversely, RetinaNet displayed a higher FPS of 16.24 and a comparatively smaller model size of 442 MB, with a shared learning rate. Notably, the YOLOv5 model stands out with exceptional efficiency, boasting an impressive FPS of 125 and a remarkably compact model size of 42 MB, aligning with the same learning rate as its counterparts. These findings underline the trade-offs between processing speed, model size, and learning rate, emphasizing YOLOv5’s notable suitability for real-time applications in the context of driver monitoring systems. Table 4 presents a comparative analysis of the performance of all three state-of-the-art models.

Table 4. Comparison of Faster R-CNN, RetinaNet, and YOLOv5 models for driver monitoring systems. Evaluation based on FPS, model size, and mAP @50%IoU

Model	Inference Speed (FPS)	Model Size	mAP@ IoU 50%
Faster R-CNN	8.56	835 MB	89.934
RetinaNet	16.24	442 MB	87.613
YOLOv5m	125	42 MB	93.6

Deployment on the Standalone Embedded Device

In order to make the detection system robust, the trained YOLOv5m model was further tested on a standalone embedded device, a mini CPU. Mini CPU devices are embedded AI computing platforms that provide high-performance, low-power computing support for deep learning models. The trained YOLOv5m model was tested and deployed on a mini CPU (the latest module in the series), which has a 4GB 5124 Volta GPU with Tensorboards. The model was running at 27 FPS (frames per second), which can be used for real-time applications. Figure 12 shows the working mechanism of a real-time advanced driver monitoring system.

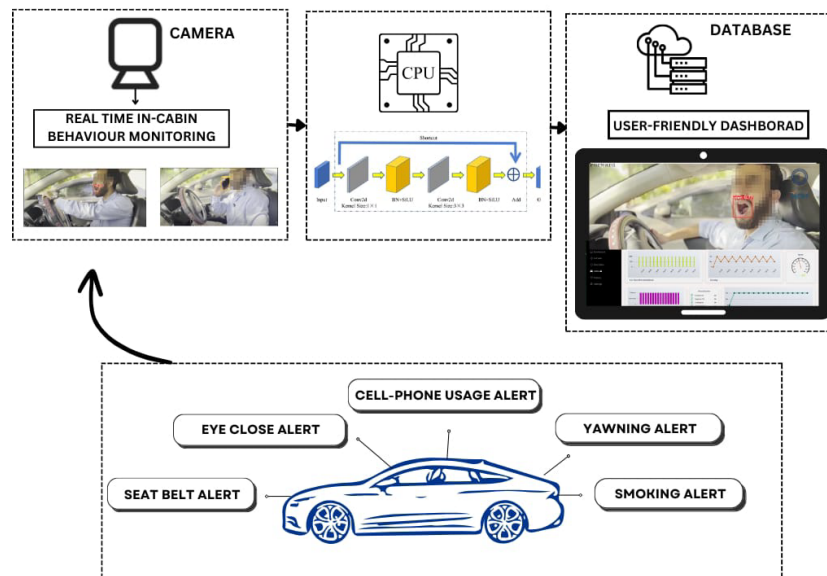


Figure 12. Functioning of advance driver monitoring system (faces in the images have been blurred to protect the privacy of individuals in accordance with ethical publishing standards).

5. Discussion

A notable future advancement involves the development of a user-friendly web application. This application will serve as a visual interface to help stakeholders, including fleet managers and safety regulators, comprehend and analyze driver behavior traits. Utilizing graphs and real-time video footage, the web application will offer an intuitive platform for monitoring and assessing driver actions. The application includes a critical feature for delivering real-time alerts and notifications to warn drivers of detected risky behaviors. This proactive capability emphasizes the system's role in enhancing road safety by facilitating immediate corrective actions. Furthermore, the application will not only support real-time surveillance but also provide historical data trends, enabling more informed decision making and effective intervention strategies.

The dataset, though diverse, may not fully capture the complete spectrum of real-world demographics and behaviors, potentially limiting the model's generalizability in edge cases. While the dataset size enabled robust training, larger datasets could further enhance performance. Future work will focus on expanding the dataset and refining collection methods to improve representation, reduce bias, and ensure broader applicability in diverse driving scenarios.

6. Conclusions

This research highlights the trade-offs between processing speed, model size, and learning rate in the context of object detection models for driver monitoring. YOLOv5m emerges as a compelling choice for real-time applications due to its outstanding efficiency and accuracy. The decision on the model to use ultimately depends on the specific requirements and constraints of the driver monitoring system, and our findings provide valuable insights for informed decision making.

Author Contributions: Conceptualization, H.Z., M.K. and F.S.; methodology, H.Z., I.u.H., M.K., N.H., F.S. and N.I.; software, I.u.H.; data curation, I.u.H., F.S. and N.I.; writing—original draft preparation, N.I.; writing—review and editing, H.Z., N.H. and F.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Abu Dhabi University's Office of Research and Sponsored Programs, grant number 19300775. The APC was funded by Abu Dhabi University.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the reason that the study involves observational data collection and analysis of driver behavior using controlled and consented data sources. It does not include any clinical interventions, sensitive personal data collection, or activities posing ethical concerns that would necessitate IRB oversight.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The dataset used in this study is publicly available and can be accessed through our research lab profile on Kaggle <https://www.kaggle.com/datasets/smartcity12/adms-dataset> (accessed on 7 October 2024).

Acknowledgments: The authors would like to thank the editors and anonymous reviewers for providing insightful suggestions and comments to improve the quality of research paper.

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

References

1. Khurshid, A.; Khan, K.; Saleem, S.F.; Cifuentes-Faura, J.; Calin, A.C. Driving towards a sustainable future: Transport sector innovation, climate change and social welfare. *J. Clean. Prod.* **2023**, *427*, 139250. [CrossRef]
2. Sehtman-Shachar, S.; Billig, P.C.; Stein, A.; Kaplan, S. The Immediate Effects of Vision-Zero Corridor Upgrades on Pedestrian Crashes in New York: A Before-and-After Spatial Point Process Approach. *Accid. Anal. Prev.* **2024**, *200*, 107531. [CrossRef] [PubMed]
3. Fink, N.J. Towards a Vision Zero Policy Theory: Examining Emerging Road Safety Initiatives in US Cities. Master's Thesis, Tufts University, Medford, MA, USA, 2016.
4. Sohail, A.; Cheema, M.A.; Ali, M.E.; Toosi, A.N.; Rakha, H.A. Data-driven approaches for road safety: A comprehensive systematic literature review. *Saf. Sci.* **2023**, *158*, 105949. [CrossRef]
5. Verma, B.; Choudhary, A. Deep learning based real-time driver emotion monitoring. In Proceedings of the 2018 IEEE International Conference on Vehicular Electronics and Safety (ICVES), Madrid, Spain, 12–14 September 2018; pp. 1–6.
6. Su, L. Vision-Based Driver State Monitoring Using Deep Learning. Master's Thesis, University of Waterloo, Waterloo, Canada, 2022.
7. Volvo Cars. Available online: <https://www.volvocars.com/mt/support/car/s60-cross-country/article/bde657a0f78c68d3c0a801e800111c9c> (accessed on 1 February 2024).
8. Mercedes-Benz of Easton. Available online: <https://www.mercedesbenzofeaston.com/mercedes-benz-attention-assist/> (accessed on 1 February 2024).
9. Choi, S.; Thalmyr, F.; Wee, D.; Weig, F. *Advanced Driver-Assistance Systems: Challenges and Opportunities Ahead*; McKinsey Company: Chicago, IL, USA; pp. 1–11. Available online: <https://www.mckinsey.com/industries/semiconductors/our-insights/advanced-driver-assistance-systems-challenges-and-opportunities-ahead> (accessed on 12 July 2024).
10. Fan, X.; Wang, F.; Song, D.; Lu, Y.; Liu, J. Gazmon: Eye gazing enabled driving behavior monitoring and prediction. *IEEE Trans. Mob. Comput.* **2019**, *20*, 1420–1433. [CrossRef]
11. Yao, Y.; Zhao, X.; Feng, X.; Rong, J. Assessment of secondary tasks based on drivers' eye-movement features. *IEEE Access* **2020**, *8*, 136108–136118. [CrossRef]
12. Ramirez, J.M.; Rodriguez, M.D.; Andrade, A.G.; Castro, L.A.; Beltran, J.; Armenta, J.S. Inferring drivers' visual focus attention through head-mounted inertial sensors. *IEEE Access* **2019**, *7*, 185422–185432. [CrossRef]
13. Zhang, G.; Etemad, A. Capsule attention for multimodal EEG-EOG representation learning with application to driver vigilance estimation. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2021**, *29*, 1138–1149. [CrossRef] [PubMed]
14. Liu, T.; Yang, Y.; Huang, G.B.; Yeo, Y.K.; Lin, Z. Driver distraction detection using semi-supervised machine learning. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 1108–1120. [CrossRef]
15. Ahmed, K.B.; Goel, B.; Bharti, P.; Chellappan, S.; Bouhorma, M. Leveraging smartphone sensors to detect distracted driving activities. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 3303–3312. [CrossRef]
16. Billah, T.; Rahman, S.M.; Ahmad, M.O.; Swamy, M.N.S. Recognizing distractions for assistive driving by tracking body parts. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 1048–1062. [CrossRef]
17. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 142–158. [CrossRef] [PubMed]
18. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
19. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1137–1149. [CrossRef] [PubMed]
20. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
21. Draz, H.U.; Khan, M.Z.; Khan, M.U.G.; Rehman, A.; Abunadi, I. A Novel Ensemble Learning Approach of Deep Learning Techniques to Monitor Distracted Driver Behaviour in Real Time. In Proceedings of the 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), Riyadh, Saudi Arabia, 6–7 April 2021; pp. 251–256.
22. Cao, C.; Wang, B.; Zhang, W.; Zeng, X.; Yan, X.; Feng, Z.; Liu, Y.; Wu, Z. An improved Faster R-CNN for small object detection. *IEEE Access* **2019**, *7*, 106838–106846. [CrossRef]
23. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
24. Diwan, T.; Anirudh, G.; Tembhurne, J.V. Object detection using YOLO: Challenges, architectural successors, datasets and applications. *Multimed. Tools Appl.* **2023**, *82*, 9243–9275. [CrossRef] [PubMed]
25. Azevedo, P.; Santos, V. YOLO-Based Object Detection and Tracking for Autonomous Vehicles Using Edge Devices. In *ROBOT2022: Proceedings of the Fifth Iberian Robotics Conference: Advances in Robotics, Zaragoza, Spain, 23–25 November 2022*; Springer International Publishing: Cham, Switzerland, 2022; Volume 1, pp. 297–308.

26. Poon, Y.S.; Lin, C.C.; Liu, Y.H.; Fan, C.P. YOLO-based deep learning design for in-cabin monitoring system with fisheye-lens camera. In Proceedings of the 2022 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 7–9 January 2022; pp. 1–4.
27. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *42*, 2980–2988.
28. Li, J.; Li, C.; Fei, S.; Ma, C.; Chen, W.; Ding, F.; Wang, Y.; Li, Y.; Shi, J.; Xiao, Z. Wheat ear recognition based on RetinaNet and transfer learning. *Sensors* **2021**, *21*, 4845. [[CrossRef](#)] [[PubMed](#)]
29. Santos, A.; Marcato Junior, J.; de Andrade Silva, J.; Pereira, R.; Matos, D.; Menezes, G.; Higa, L.; Eltner, A.; Ramos, A.P.; Osco, L.; et al. Storm-drain and manhole detection using the RetinaNet method. *Sensors* **2020**, *20*, 4450. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.