electronics

MDPI

Article

3D-CNN Method for Drowsy Driving Detection Based on Driving Pattern Recognition

Jimin Lee, Soomin Woo and Changjoo Moon *

Department of Smart Vehicle Engineering, Konkuk University, Seoul 05029, Republic of Korea; easymean0417@konkuk.ac.kr (J.L.); soominwoo@konkuk.ac.kr (S.W.)
* Correspondence: cjmoon@konkuk.ac.kr

Abstract: Drowsiness impairs drivers' concentration and reaction time, doubling the risk of car accidents. Various methods for detecting drowsy driving have been proposed that rely on facial changes. However, they have poor detection for drivers wearing a mask or sunglasses, and they do not reflect the driver's drowsiness habits. Therefore, this paper proposes a novel method to detect drowsy driving even with facial detection obstructions, such as masks or sunglasses, and regardless of the driver's different drowsiness habits, by recognizing behavioral patterns. We achieve this by constructing both normal driving and drowsy driving datasets and developing a 3D-CNN (3D Convolutional Neural Network) model reflecting the Inception structure of GoogleNet. This binary classification model classifies normal driving and drowsy driving videos. Using actual videos captured inside real vehicles, this model achieved a classification accuracy of 85% for detecting drowsy driving without facial obstructions and 75% for detecting drowsy driving when masks and sunglasses are worn. Our results demonstrate that the behavioral pattern recognition method is effective in detecting drowsy driving.

Keywords: drowsy driving; 3D-CNN; video classification

Citation: Lee, J.; Woo, S.; Moon, C. 3D-CNN Method for Drowsy Driving Detection Based on Driving Pattern Recognition. *Electronics* **2024**, *13*, 3388. https://doi.org/10.3390/electronics13173388

Academic Editors: Guanghui Yue, Wei Zhou and Wenhan Yang

Received: 5 July 2024
Revised: 16 August 2024
Accepted: 25 August 2024
Published: 26 August 2024

1. Introduction

1.1. Background

Drowsy driving impairs a driver's concentration and reaction time, leading to problems such as attention deficit and cognitive decline. For these reasons, drowsy driving doubles the risk of car accidents, threatening driver safety [1]. According to the Centers for Disease Control and Prevention (CDC), approximately 1 in 25 adult drivers have fallen asleep at the wheel [2], and the National Sleep Foundation (NSF) reports that about 6400 people die annually due to drowsy driving accidents [3].

Drowsy driving also affects logistics, as delivery operators may suffer damage due to the drowsy driving of drivers. Despite policies to maintain drivers' attention, such as mandatory breaks between stretches of driving, drivers are exposed to the risks of drowsy driving due to long driving shifts. Moreover, large cargo vehicles need longer braking distances, and the weight and size of cargo can result in large accidents with heavy casualties and financial losses.

To address this issue, monitoring technologies have been developed to detect the driver's condition in real-time. For instance, a Korean logistics company, Hanjin, has conducted a 'Smart Drowsy Driving Response System' project to prevent drowsy driving accidents among logistics workers [4]. The system detects moments when the driver is drowsy in real time, allowing for early intervention. Additionally, the Korean automotive company Hyundai Mobis has developed the "M.Brain" system, which detects drowsy driving through brain waves [5]. This system reduces the risk of accidents by alerting the driver when drowsiness is detected. To ensure safety, ideas such as warning the driver or activating autonomous driving mode when drowsy driving is detected have been

proposed, and research to determine drowsy driving states continues to advance to realize such systems.

In this study, we aim to determine drowsy driving using a camera due to the great advantage of its low cost compared to other drowsy driving detection methods. Previous studies using cameras detected drowsy driving through facial changes, but they become ineffective when the drivers wear sunglasses or face masks. To address this issue, we propose a 3D Convolutional Neural Network (3D-CNN) method that learns behavioral patterns instead of relying on facial changes. Our 3D-CNN method is inspired by the Inception structure of GoogleNet, and it performs detection effectively by learning the characteristics of various drowsy habits of individuals.

1.2. Research Proposal

The contribution of this research is as follows. First, we develop an algorithm to detect drowsy driving even for drivers wearing masks and sunglasses, using not only the driver's facial features but also the overall behavior patterns. Second, drowsy driving can be detected regardless of differences in individual drivers' drowsiness habits. By learning both drowsy behaviors and actions taken to stay awake, the system can overcome variations in drowsiness habits. Third, the algorithm demonstrates reasonable computation time by a shallow neural network framework running on an embedded PC within the vehicle. We validate the algorithm performance using data captured inside actual vehicles.

The manuscript is organized as follows. Section 2 presents the limitations of existing studies through a literature review and outlines the direction of this research. Section 3 describes the data collection and processing for training and testing the model. Section 4 proposes the learning model to detect drowsy driving. Section 5 explains our experiments and analyzes the results. We conclude our research with the implications of our results in Section 6.

2. Literature Review

2.1. Drowsy Driving Detection Methods

The two most prominent methods to detect drowsy driving are to use biometric signals or images from cameras. The biometric-signal-based methods include detecting drowsy driving through the analysis of electroencephalograms (EEG) [6–8], electrocardiograms (ECG) [9], and electromyograms (EMG) [10]. These methods involve attaching measuring devices to the driver to obtain biometric signal data, which are then used to detect drowsy driving through algorithms or deep learning models. These methods are highly accurate and capable of capturing not only the driver's drowsy state but also states of decreased attention. However, wearing measuring devices can be uncomfortable for the driver and the cost of obtaining data is high. For these reasons, biometric-signal-based drowsy driving detection has significant practical drawbacks due to its invasive nature, which can cause discomfort for the driver.

Consequently, research on drowsy driving detection has predominantly focused on camera-based methods to effectively detect drowsy driving in real vehicle environments. The camera-based drowsy driving detection method involves filming the driver's face with a camera and determining drowsy driving based on the features of the images. This approach is advantageous due to its lower cost and quick detection capabilities. A traditional method within this approach is facial landmark detection [11,12]. This method calculates the distance and movement of landmarks to detect drowsy driving by considering the frequency of eye closures, yawning, and head nodding. There are also studies [13,14] that classify drowsy images using CNN models without extracting separate landmarks. The CNN-based method classifies images of closed eyes and yawning. However, these traditional camera-based methods have limitations in that they cannot detect drowsy driving if the driver is wearing sunglasses or a mask. This is because the detection criteria, such as the eyes and mouth, cannot be recognized when covered by sunglasses or a mask, making drowsy driving detection impossible. Therefore, traditional camera-based methods

have limited applicability in terms of versatility, as they can only detect drowsy driving in situations where the driver's face is not obstructed.

To address this, research based on Transformers [15] has been conducted to demonstrate that drowsy driving can still be detected even when the driver is wearing a mask. The study in [15] utilized the attention mechanism of Transformers to bypass areas covered by the mask and focus on the uncovered facial features to detect drowsy driving. This research showed that it is possible to detect drowsy driving even when a mask is worn. However, these studies [11–15] all detected drowsy driving using only facial features, even though driver behaviors such as closing the eyes or opening the mouth can frequently occur even when not drowsy. Moreover, there may be more features beyond eye closure and yawning that can assist in detection but have not been captured in modeling. Therefore, existing studies are limited to individuals whose drowsiness habits are limited to eye blinking and yawning, which prevents them from achieving generalized performance.

Therefore, this paper aims to determine drowsy driving by recognizing the overall behavior patterns of the driver, in addition to the driver's facial features. This approach can overcome the limitations of detection accuracy due to facial obstructions, such as masks or sunglasses. We also target to learn the behavior patterns of various individuals, who may exhibit different habits when drowsy, such as covering their mouths with their hands when yawning or stretching to shake off drowsiness. This can enhance the accuracy of classifying their attention level.

Moreover, research on drowsy driving detection has faced difficulties due to the lack of datasets for drowsy and normal driving, making it challenging to conduct studies with data from actual vehicle interiors. Most drowsy driving detection studies, except for a few [12], have evaluated using everyday data or the researchers' own images rather than data from within vehicles, making it uncertain whether similar performance can be achieved in a real driving environment. Therefore, we directly collect the image data of drivers inside actual vehicles to reflect the driving environments more realistically, with drivers wearing masks and sunglasses.

2.2. Behavior Pattern Recognition

Models that classify behavior through video data often use the LSTM (Long Short-Term Memory) model, Skeleton-based model, and 3D-CNN. LSTM was introduced to overcome the problem that RNNs (Recurrent Neural Networks) are limited in learning long-term causal relationships [16]. LSTM, a type of RNN, introduces a structure of gates and memory cells to maintain both long-term and short-term memory. This allows it to process time-series data and is used to recognize behavior by utilizing temporal information in video sequences [17,18]. While LSTM excels at learning continuous movements, it may struggle with the precise recognition of individual actions. To overcome this, research on CNN (Convolutional Neural Networks) was proposed to accurately extract motion features [19–21].

Skeleton-based models recognize behavior using skeleton data, representing the positions of a person's joints. The GCN (Graph Convolutional Network) is a model frequently used to process such skeleton data. It converts skeleton data into a graph and classifies behavior based on the relationships between each node and its neighboring nodes [22,23]. However, the GCN method based on skeleton data has lower performance than video-based methods because it does not utilize detailed visual cues [23]. These methods do not consider facial expressions and miss out on important information that facial recognition can provide. Additionally, it has limitations in robustness, interoperability, and scalability, leading to the development of CNN [24].

2D-CNN (2D Convolutional Neural Network) is a deep learning model primarily used for processing image data. It automatically extracts features from images, ranging from low-level features (e.g., edges, corners) to high-level features (e.g., objects, faces), through convolution operations. 3D-CNN is an extension of 2D-CNN and has been used for action recognition [25–27]. 3D-CNN is particularly well-suited for recognizing specific

actions. Also, similar to the 2D-CNNs that effectively extract object features, 3D-CNNs perform convolution operations, making them useful for learning visual features related to actions. Due to the superior performance of the 3D-CNN, we select this method to detect the drowsiness of drivers using information that can be gathered by single images (such as closed eyes or open mouths), as well as continuous actions (such as rubbing arms and shaking heads).

The CNN model proposed in this study is built based on the Inception structure of GoogleNet [28]. GoogleNet is a 2D-CNN that performs convolution operations in parallel with 1×1, 3×3, and 5×5 filters to effectively extract feature maps. The idea is that performing operations with different kernel sizes allows the network to learn different features. Using this structure, it is possible to extract features related to relatively small changes such as eye movements and relatively large changes such as body movements, making it useful for learning various drowsiness habits. Based on this Inception structure, our model proposes a 3D-CNN model composed of two 3D convolution layers and two Inception layers. Additionally, to detect drowsy driving, the model must be executable on an embedded PC within the vehicle. Therefore, we propose a shallower neural network compared to the original GoogleNet to ensure that the model can run on an embedded PC.

3. Data

3.1. Data Collection

The data used for training and prediction were recorded inside an actual vehicle. The vehicle used is shown in Figure 1.

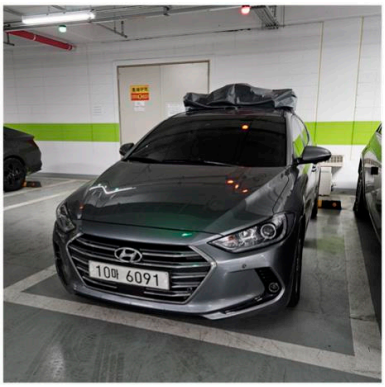

Figure 1. Hyundai Avante (mid-sized sedan) used for data collection.

The camera used for recording was the Arducam 8PM (USB Camera). The data were saved using the OpenCV library in an Ubuntu 18.04 environment, with recordings made at 10 fps and an image resolution of 1280×720. Our study does not require a high fps since it does not demand the natural continuity of actions, but rather focuses on capturing images where specific actions appear. For instance, if the purpose were to measure the frequency of eye blinking, a high fps would be necessary to ensure that no moment of the eye closing and opening is missed. However, since our study aims to learn the visual characteristics of specific actions rather than the frequency of the actions, meaningful features can be extracted even with a lower fps. The camera was installed in the center of the vehicle interior, as shown in Figure 2, to capture the driver.

The data were recorded as 5 s video clips (MP4) using the OpenCV library. OpenCV (Open Source Computer Vision) is a library that provides various functions for implementing and running image processing, computer vision, and machine learning algorithms. This study, which focuses on detecting drowsy driving, requires both normal driving data and drowsy driving data. The normal driving data were recorded in actual driving situations, while the drowsy driving data were recorded in staged scenes in a stationary situation to ensure the driver's safety. The normal driving behaviors are shown in Table 1, and the drowsy driving behaviors are shown in Table 2.

Figure 2. Arducam 8MP camera installed inside the vehicle.

Table 1. List of driver behaviors included in normal driving videos.

Class	Action
Normal Driving	Actions while driving straight Actions while stopped Actions while making a left turn Actions while making a right turn Actions while making a U-turn Actions while parking Vehicle operation Singing Talking Handling a mobile phone

Table 2. List of driver behaviors included in drowsy driving videos.

Class	Action
Drowsy Driving	Driving Yawning Nodding off Touching the neck Slapping the cheek Tapping the shoulder Gripping something Clapping Looking at something Rubbing the arms Rubbing the eyes Blinking Shaking the head side to side

The normal driving data were recorded in actual driving situations. Therefore, the driving scenarios listed in Table 1, such as "actions while driving straight" and "actions while stopped", include all typical driving behaviors like forward gaze, side gaze, and speed control. To obtain normal driving data, an alert driver must drive in actual road conditions. For this purpose, the driver operates the vehicle in various situations (such as driving straight, making right turns, etc.) in Gwangjin, Seoul, South Korea. The data collection takes approximately 13 to 35 min per driver. By collecting data in a short amount

of time, it is possible to reduce the driver's fatigue and obtain meaningful normal driving data. The normal driving data obtained in this way follow the behaviors listed in Table 1. The drowsy driving data were recorded by filming drivers acting out predetermined behaviors. Table 2 uses the list of drowsy driving behaviors from the "Monitor driver and occupant health and abnormal behavior" data provided by AIHub [29]. AIHub is a Korean website that supports various AI infrastructure and shared data. In addition to the behaviors listed in Table 2, AIHub also includes "tapping the thigh". However, this behavior was excluded from this study because the camera does not capture the driver's legs. According to Table 2, the drowsy driving data include various drowsiness habits that can occur when a driver is drowsy. Thus, unlike previous studies that judged drowsy driving based on the number of times the eyes closed or yawning, this approach can reflect a driver's various drowsiness habits. This allows the model to consistently perform well regardless of the different drowsiness habits of drivers. For example, it can apply to drivers who yawn with their mouths covered and those who yawn without covering their mouths. It can also apply to drivers who perform actions to shake off drowsiness (e.g., shaking their heads from side to side) rather than yawning. Therefore, the model can accurately detect drowsy driving regardless of the different precursor symptoms of drowsiness that drivers exhibit. This means that the model can maintain consistent performance even when the driver changes. To obtain drowsy driving data, the driver performs the behaviors listed in Table 2. Each 5 s video can include between one and three drowsy behaviors. The data collection takes approximately 13 to 35 min per driver. To reflect the habits of various drivers, no restrictions are imposed other than performing the behaviors listed in Table 2. An example of a restriction would be "not covering the mouth to ensure yawning is clearly visible." The reason for not imposing such restrictions is to enable drowsiness detection regardless of whether the driver covers their mouth while yawning or not. The drowsy driving data obtained in this way follow the behaviors listed in Table 2. The data were recorded in various weather conditions, including clear days and snowy and rainy conditions.

Figure 3 shows examples of the captured data at 1 s intervals. The first row in Figure 3 depicts a driver nodding off and yawning. The second row shows a driver rubbing their arms and tapping their shoulders, while the third row shows a driver yawning and rubbing their eyes.

Figure 3. Example of the recorded video data (the actual data used do not obscure the face).

Table 3 shows the sample sizes of various cases for model training and evaluation. According to Table 3, the training data for normal driving consist of 1890 clips: 360 clips each for four drivers, 150 clips for one driver, and 300 clips for the remaining driver. The drowsy driving data in the training set include both directly filmed videos and videos downloaded from AIHub. The directly filmed videos consist of 360 clips each for four

drivers. AIHub provides a JSON file that describes the data. The JSON file includes a category key that indicates the type of behavior in the video (e.g., drowsy driving) and an action key that specifies the behaviors included in the video (e.g., yawning, blinking, etc.). In this study, 684 videos were randomly selected and used where the category key corresponds to drowsy driving. This set involves a total of six drivers, increasing the driver diversity in the training data. Consequently, the training data comprise a total of 4014 clips, including data from 12 drivers in total (six directly filmed, six from AIHub).

Table 3. Training and testing data construction.

			Normal	Drowsy	Number of Videos	Number of Drivers
Training (No Obstruction)			1890	2124 Directly filmed: 1440 AIHub: 684	4014	12 Directly filmed: 6 AIHub: 6
Testing	No Obstruction		450	450	900	3
	Facial Obstruction	Mask	150	150	600	
		Sunglasses	150	150		

We aim to demonstrate that drowsy driving can be detected using behavior pattern recognition methods trained with the data without facial obstruction, by testing on the data with and without facial obstruction. The testing data are divided into one set with no facial obstructions and another set with facial obstructions, where masks and sunglasses are worn. For the data without obstruction, 150 normal driving and 150 drowsy driving videos were obtained from each of three drivers, totaling 900 videos. The data with facial obstruction are divided into a mask-wearing set and a sunglasses-wearing set. For both sets, 50 normal driving and 50 drowsy driving videos were obtained from each of the three drivers. The data with obstruction therefore totaled 600 videos from three drivers. The drivers from the testing data do not overlap with the drivers from the training data.

3.2. Data Preprocessing

The original captured image has a size of 1280 pixels by 720 pixels. Since this study focuses on learning the driver's behavior patterns, the background area, which excludes the driver, contains unnecessary information. Therefore, the background area is partially removed using OpenCV [30]. The image with the background removed is shown in Figure 4.

The top image in Figure 4 is the original image, and the bottom image is the image with the background partially removed using OpenCV. The size of the image with the background removed is 880 × 720. The image with the background removed is resized to 224 × 224 pixels for computational efficiency in the 3D-CNN model. A 5 s video is processed at three frames per second, resulting in a total of 15 frames. Tensorizing these frames provides the final input value for the deep learning model. We convert the video data format to a tabular data format, tagged by a label. We use the label value 0 for normal driving and 1 for drowsy driving. Therefore, the videos corresponding to "Normal" in Table 3 are assigned a label of 0, and the videos corresponding to "Drowsy" are assigned a label of 1. In other words, all videos of alert drivers actually driving (following the behaviors in Table 1) are labeled as 0, while all videos of drivers demonstrating the behaviors in Table 2 while stationary are labeled as 1. Normal driving and drowsy driving are recorded in separate situations, so the labels can be assigned automatically without any additional classification process.

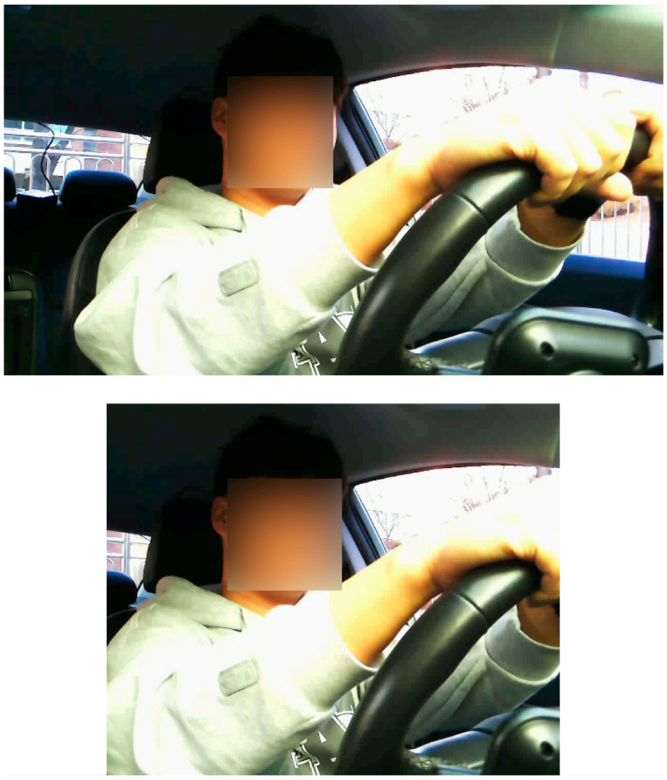

Figure 4. The image with the background removed using OpenCV (the actual data used do not obscure the face).

4. Drowsy Driving Detection Model

In this section, we propose the architecture of the 3D-CNN model for detecting drowsy driving with facial changes and behavioral action and describe how this model is trained.

4.1. 3D-CNN Model with GoogleNet Inception Structure

The Inception structure performs convolution operations in parallel with multiple filters [28]. We use this method and perform GoogleNet convolution operations in parallel with 1×1, 3×3, and 5×5 filters to effectively extract feature maps. This means that performing convolution operations with kernels of various sizes allows for the extraction of features with different scales.

The Inception structure has been used not only for image processing but also in research for processing videos. I3D (Two-Stream Inflated 3D ConvNet) is a representative example that significantly improved action classification performance using a 3D-CNN approach reflecting the Inception structure [31]. Our paper proposes a 3D convolutional architecture based on the Inception structure of GoogleNet, as shown in Figure 5. The model in Figure 5 performs convolution operations in parallel using 1×1, 3×3, and 5×5 filters, similar to the Inception structure. To detect drowsy driving, it is necessary to reflect not only small-area features such as blinking or yawning, but also large-area features such as slapping the cheek or tapping the shoulder. Therefore, various features are learned through filters of different sizes to capture both small and large changes.

This study constructs the model with 3D convolution layers, unlike GoogleNet, because it takes videos as input. Drowsy driving is limited to 13 behaviors, as shown in Table 2, and there is a possibility of overfitting due to the use of data recorded in staged scenes in a stationary situation. Therefore, after passing through two 3D convolution layers, an auxiliary classifier that passes through a linear classifier was designed. The intermediate result corresponds to the Aux result in Figure 5. The auxiliary classifier helps reduce the possibility of overfitting and enhances the generalization performance of the model. The auxiliary classifier is used during the training process but not during the prediction

process. The final architecture consists of two 3D convolution layers and two Inception structures. The reason for designing the model with two 3D convolutional layers and two Inception layers is to reduce the possibility of overfitting, as mentioned earlier. The training data obtained in this study include only 12 drivers, not hundreds. If the model were composed of very deep layers, it might show performance tailored specifically to those 12 drivers rather than achieving generalized performance. Additionally, since the drowsy behaviors are limited to 13 actions, a model with very deep layers could overfit and incorrectly classify normal driving as drowsy driving. Therefore, to avoid overfitting on the limited amount of data and the constrained features of drowsy behaviors, the model was designed with a relatively shallow structure, consisting of two 3D convolutional layers and two Inception layers.

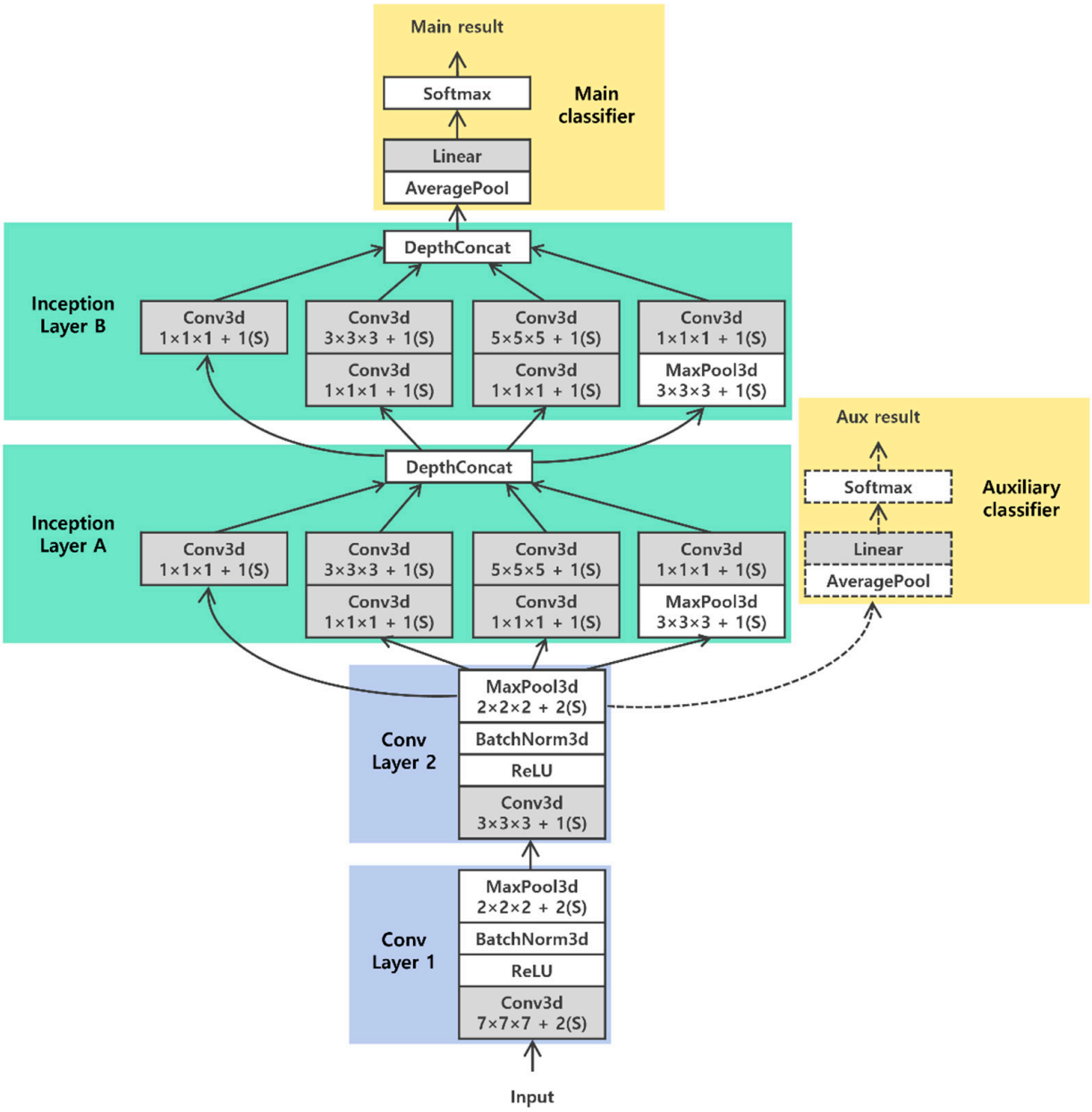

Figure 5. The drowsy driving detection model architecture proposed in this paper.

As shown in Figure 5, the input first passes through two convolutional layers. These layers perform convolution operations with a single kernel size, allowing the extraction of the overall key features from the input. Next, the input passes through two Inception layers,

where convolution operations are performed with multiple kernel sizes. This enables the model to learn features related to both small and large drowsy behaviors. The auxiliary classifier operates after the input passes through the two convolutional layers, while the main classifier operates after the input passes through the two convolutional layers and the two Inception layers. Both the auxiliary and main classifiers are used for loss calculation. Considering that the model in Figure 5 will be executed on an embedded PC in the vehicle in future research, it was built as a shallower neural network than the original GoogleNet. After passing through the last Inception layer, average pooling was used to prevent network overfitting, and the Main result, corresponding to the final output, was converted into a probability distribution through the softmax function. The reason for setting the activation function to softmax is to enhance the stability of the training process by using the cross-entropy function, a loss function, during training. This approach helps guide the model to learn gradually, especially in ambiguous boundary cases, without causing significant distortions in the wrong direction. All convolution operations maintain the spatial size of the feature map through appropriate pooling and stride. The design of other layers follows the rules of GoogleNet.

Table 4 is an architecture that shows the output of each layer. According to Table 4, the output of the convolution operations with 3×3 and 5×5 kernel sizes, which are suitable for reflecting the drivers' behavior patterns, is 128 channels. The 1×1 convolution operations, which are suitable for learning facial features, output 16 and 32 channels, respectively.

Table 4. Model architecture specifications (supplementary to Figure 5).

	Output	# 1×1 Output	# 3×3 First	# 3×3 Output	# 5×5 First	# 5×5 Output	Pool Output	In_Features	Out_Features
Conv Layer 1	$56 \times 56 \times 32$								
Conv Layer 2	$28 \times 28 \times 64$								
Inception Layer A	$28 \times 28 \times 64$	16	32	128	32	128	64		
Inception Layer B	$28 \times 28 \times 128$	32	32	128	32	128	64		
Auxiliary classifier								64	2
Main classifier								128	2

4.2. Training Method

The maximum number of epochs for training is 10, and the batch size is 4. The Adam optimizer with a learning rate of 0.0003 was used. The loss function is cross-entropy. Let l_{AC} be the loss for Aux result of the auxiliary classifier in Figure 5 and l_M be the loss for Main result of the main classifier. The l_{total} follows example 1 of an equation:

$$l_{total} = (1 - \alpha) \cdot l_M + \alpha \cdot l_{AC} \quad (1)$$

The values of α experimented with were 0.2, 0.3, 0.4, and 0.5. While the differences were not significant, the best performance was observed when training with α set to 0.2. Additionally, the model with the auxiliary classifier showed an average improvement of about 2% in AUC compared to the model without the auxiliary classifier. The training was conducted in a Jupyter notebook environment using a single NVIDIA RTX A6000 GPU (NVIDIA, Santa Clara, CA, USA) and took approximately 80 min for 10 epochs.

5. Experiments

In this chapter, we will analyze the results to evaluate the classification performance and computation time of the model.

5.1. Prediction Accuracy

To evaluate the classification performance of the model, a cross-validation approach was employed. Evaluating the model using cross-validation can reduce bias and assess the variability of the model's performance. This allows for an evaluation of the model's reliability and stability. The cross-validation process is as follows: first, 80% of the entire training data from Table 3 is randomly selected, and training is conducted using the method described in Section 4.2. Note that the 80% of data was selected for each category, normal and drowsy, so that the training data are not skewed towards any one category. Once the training is completed, 80% of the testing data from Table 3 is selected for prediction. In other words, 80% of the data is sampled and selected for each type: with and without facial obstruction, and with masks vs. sunglasses. This process is repeated a total of 10 times for cross-validation. The accuracy is calculated based on the prediction results in terms of the average and standard deviation over these 10 iterations.

Figure 6 depicts the average accuracy (indicated on the bars) and standard deviations (indicated at the ends of the bars). We see that without facial obstruction, the average accuracy is 0.679 for normal driving and 0.854 for drowsy driving. The high accuracy score for drowsy driving validates that the model can accurately detect dangerous situations when drowsy driving occurs and prevent traffic accidents in advance. The average accuracy for normal driving is somewhat lower than for drowsy driving; however, this can be interpreted as the algorithm making conservative predictions on drowsy driving.

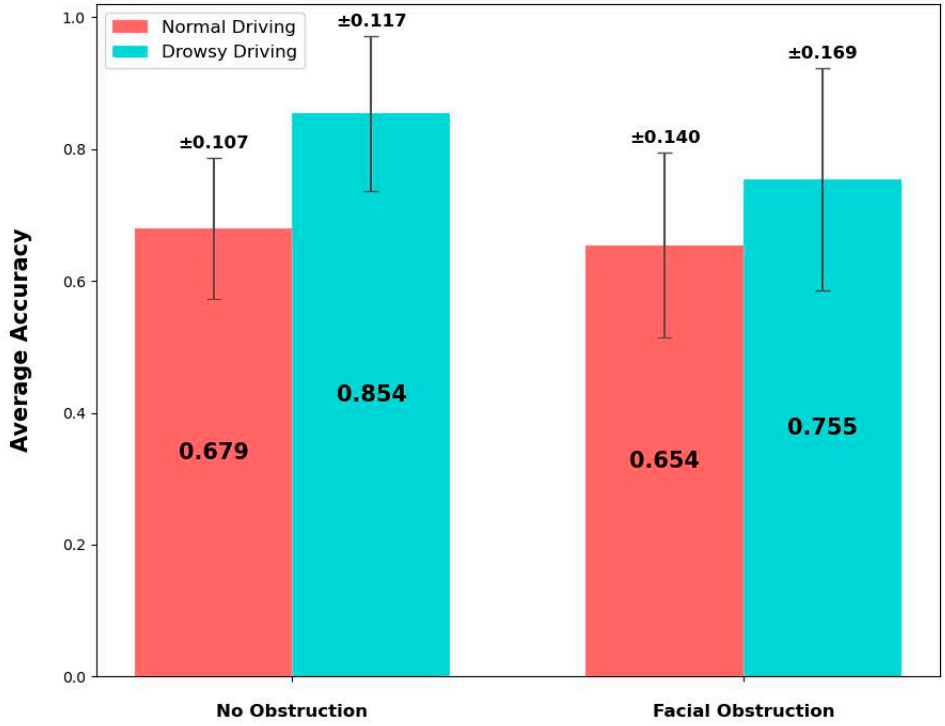

Figure 6. Accuracy of the test data.

On the right of Figure 6, the average accuracy with facial obstruction is 0.654 for normal driving and 0.755 for drowsy driving. Compared to the cases without obstruction, the average accuracy for normal driving is not significantly different, but the average accuracy for drowsy driving has decreased by about 0.1. The habits that best represent drowsiness are "blinking" and "yawning". Analysis reveals that the accuracy decreased by about 0.1 because wearing a mask and sunglasses covers these most representative signs of drowsiness. Additionally, when the face is covered, subtle facial expression changes that occur when drowsy cannot be reflected. This interference with the availability of facial information is interpreted as having affected the model's performance. Still, the reasonable accuracy of around 0.755 exhibits that our algorithm with behavior pattern

recognition can overcome the limitations of existing research that uses only facial changes and cannot detect drowsy driving with masks and sunglasses worn. Our goal is to achieve reliable accuracy through behavior pattern recognition, even if we cannot detect the most representative drowsy behaviors, such as blinking and yawning. Therefore, we have demonstrated that even in situations where blinking or yawning cannot be detected, we can still detect drowsiness with an accuracy of approximately 0.755 by recognizing other drowsy behavior patterns. This demonstrates that behavior pattern recognition can detect drowsy driving even with facial obstruction.

Moreover, the standard deviations for the cases without and with facial obstruction are approximately 0.1 for both normal and drowsy driving. This indicates that the model performance is relatively consistent for all cases, demonstrating enhanced reliability of the accuracy.

In Figure 7, we present the average accuracy (indicated on the bars) and standard deviations (indicated at the ends of the bars) for three drivers that were not in the training dataset without facial obstructions. The average accuracies range around 0.75 to 0.78, with standard deviations ranging from 0.065 to 0.155. This confirms that the model maintains consistent performance regardless of the driver. Recent AI-based studies like [13–15] could not secure a sufficient amount of data per driver due to the issue of data scarcity, which prevented them from performing driver-specific accuracy evaluations. As a result, the reliability of these methods in maintaining consistent performance when the driver changes cannot be assured. However, our study demonstrated that our approach and training methodology can consistently achieve stable model performance by measuring the accuracy for each individual driver.

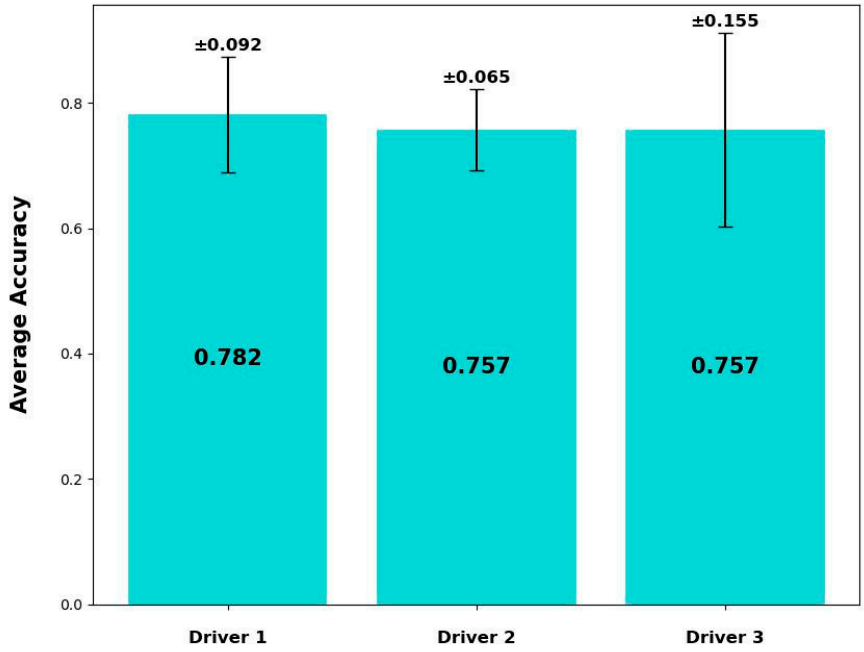

Figure 7. Accuracy for new drivers without facial obstructions.

We next dive deeper into the accuracy of facial obstruction cases. Figure 8 shows the average accuracy (indicated on the bars) and standard deviations (indicated at the ends of the bars) for the facial obstruction test cases, divided into a mask-wearing set and a sunglasses-wearing set. These two sets are intrinsically different because these items cover the faces in different areas; the masks cover the nose, mouth, and chin, whereas the sunglasses cover the eye movements. The average accuracy is about 0.1 lower when wearing sunglasses compared to wearing a mask. This indicates that the model accuracy is impacted by the changes in the eyes more significantly than the changes in the mouth for detecting drowsy driving (assuming that the nose does not exhibit much information).

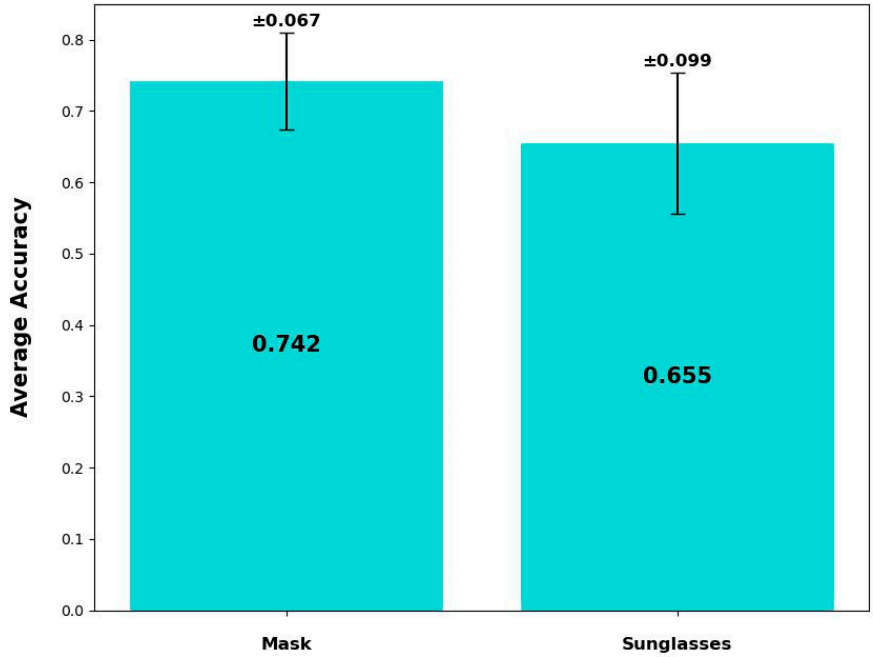

Figure 8. Accuracy with facial obstruction, divided into groups with masks and sunglasses.

5.2. Computation Time

We now evaluate the computation time of prediction from 1500 samples. The computation time to determine whether a 5 s video clip shows drowsy driving or normal driving is approximately 0.1 s. This rapid prediction time is likely due to the model being built as a shallow neural network. On average, drivers take between 0.9 and 1.5 s to apply the brakes in unexpected situations [32]. Therefore, our model's computation time of 0.1 s is expected to provide an appropriate warning or activation of safety measures in a safe timeframe. Thus, the prediction time of this model is sufficiently fast for use in real-world situations.

Recent studies in the field of Computer Vision, such as [15], have seen a surge in research focusing on image classification using attention mechanisms [33]. For instance, [15] demonstrated the effectiveness of detecting drowsiness in data where masks are worn, leveraging attention mechanisms. However, attention mechanisms face the challenge of increased computational time due to the need to calculate relationships between every position in the input and every other position. To address this issue, methods like Swin Transformer [34] and Convolutional Vision Transformer [35] have been proposed to reduce computational load. Nonetheless, these approaches still face difficulties when applied to tasks requiring real-time processing. Therefore, our model's ability to achieve both high accuracy and fast computation time demonstrates the viability of the 3D-CNN approach in real-world scenarios where rapid detection of drowsy driving is critical.

6. Conclusions

This study proposes a camera-based method for detecting drowsy driving, by recognizing the driver's facial features and actions, that overcomes the limited detection accuracy of methods in the current literature when facial obstructions (such as masks or sunglasses) are worn by drivers. The training and test data used were directly captured inside actual vehicles, with additional data from the AIHub. We develop a 3D-CNN based on the GoogleNet Inception structure for binary classification of video data.

From cross-validation, we demonstrate the model's performance in detecting drowsy driving with a classification accuracy of 85% for drowsy driving. Additionally, the model achieved an accuracy of 75% for drowsy driving even when the driver was wearing a mask and sunglasses. To the best of our knowledge, our approach overcomes the critical limitation of the current literature's proposed methods, which fail to detect drowsy driving when masks or sunglasses are worn as they only learn facial features [36]. Furthermore,

the prediction time for a 5 s video averaged at around 0.1 s, demonstrating sufficiently fast performance for real-time use. With this algorithm implemented in real life, we expect to reduce unnecessary warnings to drivers and provide appropriate measures to prevent accidents from drowsy driving.

Our work can be further developed as follows. First, to ensure the safety of the drivers in the data, the drowsy driving video data constructed in this study were created through staged scenarios rather than being collected in actual driving environments with actual drowsy driving. Therefore, further validation is needed to ensure sufficient accuracy in real-world environments.

Second, it is necessary to verify through experiments that the proposed 3D-CNN model can run on an embedded PC within a vehicle, confirming its optimization for the in-vehicle environment. This will demonstrate the model's applicability in real-world scenarios. Thirdly, a system should be developed to process incoming image data in real time, rather than classifying pre-recorded video data. This will enable the development of a real-time drowsy driving detection monitoring system.

Third, it is crucial to secure high-quality training data. The training data used in this study consist of approximately 4000 samples, which is a relatively small dataset. Moreover, the training data include a total of 12 drivers, so it is necessary to consider a wider variety of drivers. Analyzing and incorporating common drowsy habits into the training data is also expected to yield better accuracy.

By incorporating this method into a monitoring system for detecting drowsy driving, it is possible to prevent accidents caused by drowsy driving. For example, monitoring delivery drivers (or large-bus drivers) can help prevent large-scale traffic accidents. Accurate detection of drowsy driving can be utilized in alert systems that provide warning sounds or vibrations to the driver. Additionally, a service that activates autonomous driving mode upon detecting drowsy driving can also be provided.

Author Contributions: Conceptualization, J.L. and C.M.; methodology, J.L.; software, J.L.; validation, J.L.; formal analysis, J.L. and S.W.; investigation, J.L. and S.W.; resources, C.M.; data curation, J.L.; writing—original draft preparation, J.L. and S.W.; writing—review and editing, J.L., S.W., and C.M.; visualization, J.L.; supervision, S.W. and C.M.; project administration, J.L.; funding acquisition, C.M. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was supported by the Korea Institute for Advancement of Technology (KIAT) grant funded by the Korean Government (MOTIE) (P0020536, HRD Program for Industrial Innovation).

Data Availability Statement: Data are available upon request from the authors.

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

References

1. Bioulac, S.; Micoulaud-Franchi, J.-A.; Arnaud, M.; Sagaspe, P.; Moore, N.; Salvo, F.; Philip, P. Risk of Motor Vehicle Accidents Related to Sleepiness at the Wheel: A Systematic Review and Meta-Analysis. *Sleep* **2017**, *40*, zsx134. [CrossRef] [PubMed]
2. Wheaton, A.G.; Shults, R.A.; Chapman, D.P.; Ford, E.S.; Croft, J.B.; Division of Population Health, National Center for Chronic Disease Prevention and Health Promotion. Drowsy driving and risk behaviors—10 States and Puerto Rico, 2011–2012. *MMWR Morb. Mortal. Wkly. Rep.* **2014**, *63*, 557–562. [PubMed] [PubMed Central]
3. National Sleep Foundation. Drowsy Driving Survey. 2023. Available online: https://www.thensf.org/wp-content/uploads/2023/11/NSF-2023-Drowsy-Driving-Survey-Report.pdf (accessed on 24 August 2024).
4. HANJIN. 2023 ESG Report. p. 31. Available online: https://www.hanjinkal.co.kr/common/file/2023%EB%85%84%20(%EC%A3%BC)%ED%95%9C%EC%A7%84%20ESG%20%EB%B3%B4%EA%B3%A0%EC%84%9C%20(%EC%98%81%EB%AC%B8).PDF (accessed on 24 August 2024).
5. "A Brainwave Technology from Hyundai Mobis Proven to Reduce Drowsiness and Inattentive Driving by Up to 1/3", HYUNDAI MOBIS. April 2022. Available online: https://www.mobis.co.kr/en/aboutus/press.do?category=press&idx=5595 (accessed on 24 August 2024).
6. Lin, C.T.; Wu, R.C.; Liang, S.F.; Chao, W.H.; Chen, Y.J.; Jung, T.P. EEG-based drowsiness estimation for safety driving using independent component analysis. *IEEE Trans. Circuits Syst. I Regul. Pap.* **2005**, *52*, 2726–2738. [CrossRef]
7. Xu, Y.; Wu, D. EEG-Based Driver Drowsiness Estimation Using Self-Paced Learning with Label Diversity. In Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China, 6–9 December 2019; pp. 369–375. [CrossRef]

8. Li, M.A.; Zhang, C.; Yang, J.F. An EEG-based method for detecting drowsy driving state. In Proceedings of the 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, Yantai, China, 10–12 August 2010; pp. 2164–2167. [CrossRef]
9. Jahn, G.; Oehme, A.; Krems, J.F.; Gelau, C. Peripheral detection as a workload measure in driving: Effects of traffic complexity and route guidance system use in a driving study. *Transp. Res. Part F Traffic Psychol. Behav.* **2005**, *8*, 255–275. [CrossRef]
10. Akin, M.; Kurt, M.B.; Sezgin, N.; Bayram, M. Estimating vigilance level by using EEG and EMG signals. *Neural Comput. Appl.* **2008**, *17*, 227–236. [CrossRef]
11. Salzillo, G.; Natale, C.; Fioccola, G.B.; Landolfi, E. Evaluation of Driver Drowsiness based on Real-Time Face Analysis. In Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 11–14 October 2020; pp. 328–335. [CrossRef]
12. Peddarapu, R.K.; Likhita, B.; Monika, D.; Paruchuru, S.P.; Kompella, S.L. Raspberry Pi-Based Driver Drowsiness Detection. In Proceedings of the 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT), Greater Noida, India, 9–10 February 2024; pp. 864–869. [CrossRef]
13. Macalisang, J.R.; Alon, A.S.; Jardiniano, M.F.; Evangelista, D.C.P.; Castro, J.C.; Tria, M.L. Drive-Awake: A YOLOv3 Machine Vision Inference Approach of Eyes Closure for Drowsy Driving Detection. In Proceedings of the 2021 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), Kota Kinabalu, Malaysia, 13–15 September 2021; pp. 1–5. [CrossRef]
14. Mirabdullayev, I.; Ayoobkhan, M.U.A.; Hashana, A.J.; Ali, L.A.K.S. Drowsy Driving Detection System Using Face Detection. In Proceedings of the 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 1–3 November 2023; pp. 779–784. [CrossRef]
15. Qu, S.; Gao, Z.; Wu, X.; Qiu, Y. Multi-Attention Fusion Drowsy Driving Detection Model. *arXiv* **2023**, arXiv:2312.17052.
16. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
17. Shun, C.C.; bin Ibrahim, M.Z.; Muhamad, I.H.B. Human Behaviors Classification Using Deep Learning Technique. In Proceedings of the 6th International Conference on Electrical, Control and Computer Engineering: InECCE2021, Kuantan, Pahang, Malaysia, 23 August 2021; Springer: Singapore, 2022.
18. Zhao, Y.; Yang, R.; Chevalier, G.; Xu, X.; Zhang, Z. Deep residual bidir-LSTM for human activity recognition using wearable sensors. *Math. Probl. Eng.* **2018**, *2018*, 7316954. [CrossRef]
19. Baccouche, M.; Mamalet, F.; Wolf, C.; Garcia, C.; Baskurt, A. Sequential deep learning for human action recognition. In Proceedings of the Human Behavior Understanding: Second International Workshop, HBU 2011, Amsterdam, The Netherlands, 16 November 2011; Proceedings 2. Springer: Berlin/Heidelberg, Germany, 2011.
20. Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–15 June 2015; pp. 4694–4702.
21. Ibrahim, M.S.; Muralidharan, S.; Deng, Z.; Vahdat, A.; Mori, G. A hierarchical deep temporal model for group activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1971–1980.
22. Liu, Z.; Xie, Q.; Lu, Y.; Wang, X. Skeleton-based Action Recognition with Two-Branch Graph Convolutional Networks. In Proceedings of the Journal of Physics: Conference Series, Changsha, China, 13–16 August 2021; IOP Publishing: Philadelphia, PA, USA, 2021; Volume 2030.
23. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
24. Duan, H.; Zhao, Y.; Chen, K.; Lin, D.; Dai, B. Revisiting skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2969–2978.
25. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 221–231. [CrossRef] [PubMed]
26. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6450–6459.
27. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
28. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
29. AIHub. Monitor Driver and Occupant Health and Abnormal Behavior. Available online: https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=651 (accessed on 4 July 2024).
30. OpenCV Team. OpenCV Library. Version 4.6.0. 2023. Available online: https://opencv.org/ (accessed on 4 July 2024).
31. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
32. Johansson, G.; Rumar, K. Drivers' brake reaction times. *Hum. Factors* **1971**, *13*, 23–27. [CrossRef] [PubMed]
33. Vaswani, A. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.

34. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
35. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 22–31.
36. Safarov, F.; Akhmedov, F.; Abdusalomov, A.B.; Nasimov, R.; Cho, Y.I. Real-time deep learning-based drowsiness detection: Leveraging computer-vision and eye-blink analyses for enhanced road safety. *Sensors* **2023**, *23*, 6459. [CrossRef] [PubMed]