

## RESEARCH ARTICLE

# A Real-Time Vision Transformers-Based System for Enhanced Driver Drowsiness Detection and Vehicle Safety

ANWAR JARNDAL<sup>1</sup> , (Senior Member, IEEE), HISSAM TAWFIK<sup>1</sup>,  
ALI I. SIAM<sup>2,3</sup>, IMAD ALSYUOF<sup>4</sup> , AND ALI CHEAITOU<sup>4</sup>

<sup>1</sup>Department of Electrical Engineering, University of Sharjah, Sharjah, United Arab Emirates

<sup>2</sup>Research Institute of Sciences and Engineering, University of Sharjah, Sharjah, United Arab Emirates

<sup>3</sup>Department of Embedded Network Systems Technology, Faculty of Artificial Intelligence, Kafrelsheikh University, Kafr El-Sheikh 33516, Egypt

<sup>4</sup>Department of Industrial Engineering and Engineering Management, University of Sharjah, Sharjah, United Arab Emirates

Corresponding author: Anwar Jarndal (ajarndal@sharjah.ac.ae)

This work was supported by the University of Sharjah, Sharjah, United Arab Emirates.


**ABSTRACT** Drowsy driving is a leading cause of fatal traffic accidents worldwide. Drowsy driving has emerged from modern societal trends such as long working hours, heavy reliance on vehicles, and insufficient sleep. Despite considerable efforts by researchers to develop efficient driver drowsiness detection systems, none so far has been widely adopted due to their high cost, intrusive nature, and ineffectiveness in challenging real-life situations. This paper presents a novel, real-time, non-intrusive, and cost-effective driver drowsiness detection system leveraging vision transformers (ViT). Our approach detects the driver's face from each video frame and classifies the driver's state as either 'drowsy' or 'alert' based on the entire facial image, as opposed to previous systems that rely on analyzing specific facial features. We demonstrate that the proposed Vision Transformers-based Driver Drowsiness Detection (ViT-DDD) system surpasses existing state-of-the-art methods, particularly in challenging scenarios such as drivers wearing glasses or sunglasses, or in different lighting conditions. The model was trained and evaluated on two widely used public drowsiness detection datasets, achieving classification accuracies of 98.89% on the NTHU-DDD dataset and 99.4% on the UTA-RLDD dataset. Furthermore, the system was successfully deployed on a Raspberry Pi microcomputer, integrated with an infrared camera, a GSM/GPS module, and a buzzer to alert the driver and report the drowsiness condition to the vehicle owner. Testing the prototype yielded highly promising results, with the system's strong performance attributed to the ViT-DDD system and advanced hardware. The promising test results suggest the potential of this system in significantly reducing accidents caused by drowsy driving, with future work aiming to expand its capabilities and integration into broader vehicular systems.

**INDEX TERMS** Vision transformers, driver drowsiness, deep learning, computer vision, vehicle safety.

## I. INTRODUCTION

Driving is an integral part of our modern society and the world's economy. People all over the world either drive their vehicles to navigate their busy lives or rely on public buses, trains, trams, and metros. In addition to getting people to

where they want to go, driving is an essential part of the world trading network. Every day, trucks and freight trains distribute millions of goods all over the world. In the United States, truck drivers can be behind the wheel for up to 11 straight hours [1]. These long driving hours, combined with a lack of sufficient rest for most adults, as the World Economic Forum reports that 62% of people do not get sufficient sleep [2], lead to potential life-and-death problems

The associate editor coordinating the review of this manuscript and approving it for publication was Mostafa M. Fouda .

of drowsy driving. Drowsy driving is characterized by the situation where the driver operates for a part of the journey by sleepy or fatigued states while driving. Studies indicate that around 1 in 25 adult drivers in the United States has reported falling asleep while driving in any given month [3], [4]. The US National Highway Traffic Safety Administration (NHTSA) reports that sleepy driving contributed to 91,000 crashes in 2017 that resulted in 50,000 injuries [5]. The organization estimates that over 6,000 fatal accidents occur annually due to sleepy driving [6], [7]. Drowsy driving has emerged as a critical issue for highway and road safety due to its possibly fatal consequences. Many researchers have created several detecting devices to notify drivers when they are in danger. Numerous traditional driver drowsiness detection systems rely on physiological signals, including electroencephalography (EEG), electrooculography (EOG), and heart rate monitoring [8]. Although these technologies prove effective in controlled settings, their adoption in real-world scenarios has been constrained by high cost, intrusive designs, and difficulties in deployment in different driving circumstances.

Recently, advancements in computer vision and machine learning (ML) technologies have emerged new drowsiness detection systems that focus on analyzing facial features by tracking specific facial indicators such as eye closure, yawning, and head movements to assess the driver's level of alertness. However, the reliability of such systems remains limited, specifically in challenging conditions, such as when the driver wears glasses or sunglasses, or when lighting conditions vary significantly [9].

This paper presents a real-time driver drowsiness detection and alertness system based on the state-of-the-art Vision Transformers (ViT) technique. Unlike previous drowsiness detection systems that rely on analyzing specific facial features, our approach utilizes the entire face image to classify the driver's state as either 'drowsy' or 'alert.' This approach enhances the robustness of the system, allowing it to perform well even in challenging scenarios, contrary to other methods in the literature that do not have sufficient efficiency in challenging conditions. The proposed Vision Transformers-based Driver Drowsiness Detection (ViT-DDD) system was trained on two widely adopted datasets for drowsiness detection: the National Tsing Hua University Driver Drowsiness Detection (NTHU-DDD) dataset and the University of Texas at Arlington Real-Life Drowsiness Dataset (UTA-RLDD) dataset. These datasets contain real driving sessions with different driving scenarios and conditions, providing a generalization for the proposed system even with different challenging scenarios. Furthermore, the proposed ViT-DDD system was successfully deployed on a Raspberry Pi 4B board integrated with an infrared camera to provide a stand-alone solution that can be installed and used in any car.

The rest of this paper is structured as follows. Section II presents a review of related works, discussing their advantages and limitations. Section III presents the proposed

approach, starting with a technical background on the computer vision methods that form the foundation of our system, and then we introduce the structure and components of the ViT-DDD system. Section IV details the datasets used for training and evaluating our model, while Section V presents our results and findings. Section VI covers the implementation and testing of the system. Finally, Section VII concludes the paper and discusses potential directions for future work.

## II. RELATED WORK

Due to the critical nature of the drowsy driving problem, numerous researchers from both academia and the automotive industry have proposed systems to alert drivers when they become drowsy or start falling asleep while driving. These systems can be broadly categorized into three approaches based on the features used to detect drowsiness: physiological-based, vehicular-based, and behavioral-based systems.

### A. PSYCHOLOGICAL-BASED APPROACHES

In physiological-based systems, researchers have used different physiological measurements related to the activities of the central and autonomous nervous systems of the driver. Commonly used measurements include Heart Rate Variability (HRV) [10], Electroencephalogram (EEG) [11], [12], [13], Electrooculogram (EOG) [14], Electrocardiogram (ECG) [15], and neuromuscular activities from Electromyogram (EMG) [16], [17], to distinguish the drowsy state and alert the driver. The accuracy obtained in [13] was 94.31%, which is higher than the accuracies reported in [16] and [17]. This could be attributed to the hybrid technique employed using EEG signals and their spectrograms. The physiological-based systems require sensors to be attached to the driver's body which makes them uncomfortable to use over long periods and impractical for real-life applications.

### B. VEHICULAR-BASED APPROACHES

Vehicular-based drowsiness detection systems rely on sensors attached to the car that measure vehicular based attributes such as steering angle, lane departure, pedal movement, braking, steering movement, and speed. After building a profile of the personal driving habits of the driver in the first few minutes of a trip, these systems can detect irregularities in driving behavior and alert the driver to take a break. These vehicular-based systems have been utilized by several car manufacturers, such as Ford, Mercedes-Benz, and Volkswagen [18], [19]. Vehicle-based systems have reported accuracy rates of up to 90% [16]. However, due to their dependency on expensive sensors, these vehicular-based driver drowsiness detection systems are only available in high-end cars, which makes them inaccessible to a wide range of drivers [20].

### C. BEHAVIORAL-BASED APPROACHES

Due to the high-cost limitation of vehicular-based systems and the intrusive nature of physiological-based systems,

significant research efforts have focused on developing behavioral-based drowsiness detection systems [21], [22], [23]. Behavioral-based systems identify drowsiness by analyzing changes in a driver's facial features, such as Eye Aspect Ratio (EAR) [24], eye blink rate, Percentage of Eye Closure (PERCLOS), and yawning [25]. Typically, these systems extract the driver's face from a video feed, preprocess the video frames, and then feed them to a trained machine learning or deep learning classifier such as Support Vector Machines (SVMs) or Convolutional Neural Networks (CNNs). These classifiers are trained using large, labeled datasets; thus, an integral part of developing effective behavioral-based systems for drowsiness detection is the availability of such large, labeled, and relevant datasets. Many earlier behavioral-based systems relied on private datasets [26], [27], [28], [29], [30], which limits the ability to reproduce the results, and most importantly makes it hard to consistently or fairly compare different drowsiness detection methods. Therefore, we will only focus on discussing previous methods that used publicly available datasets for training and evaluation so that we can later compare our results to those other methods.

In [31], an ensemble of four CNN models was used to detect drowsiness based on four different features: hand gestures, facial expressions, behavioral features, and head movements. They reported a classification accuracy of 85% on the NTHU-DDD dataset. Using the same dataset, the work reported in [32] showed an accuracy of 81% based on a simple MLP classifier, and its accuracy was later improved by the same authors using a CNN model [33]. In [34], the authors used the blinking rate of a subject's eye as a facial feature to detect drowsiness. They used a Hierarchical Multiscale Long Short-Term Memory (HMLSTM) module for classification. They reported an accuracy of 65.2% on the RLDD dataset. In [35], Weng et al. presented a Hierarchical Temporal Deep Belief Network (HTDBN)-based method to detect driver drowsiness. Their approach starts with preprocessing inputs from the face or head and then passing them to deep belief networks for extracting high-level features. The output from the deep belief networks is an observable vector sequence, which is then analyzed using two Hidden Markov Models (HMMs) to predict the driver's drowsiness level. Finally, an inverse logit transform is used to forecast the driver's level of drowsiness [35]. They report an average classification accuracy of 84.82% on the NTHU-DDD dataset. In [36], the authors presented a driver drowsiness detection system based on a combination of CNNs and Haar cascade classifiers. Their system uses Haar feature-based cascade classifiers on an input image to extract the regions of interest from the driver's face. These regions are then fed into three separate CNNs to extract features and predict a class for each region based on defined classes. Then a final classification is obtained based on the combined prediction of each of the separate CNN models. They report an 89% classification accuracy on the RLDD dataset. In [37], the authors utilized 3D Deep Neural

Networks (DNN) to capture temporal information along with spatial information of the facial features. The architecture is intended for identifying faces and classifying a driver's drowsiness. To identify potential face areas and ensure consistency between frames, the first part of the design includes a Haar feature-based face detector. The second section employs a 3D CNN network to extract features from spatial and temporal domains. The network is made up of six convolutional layers and two fully connected layers to process a 20-frame sequence. To minimize spatial dimensionality and capture significant characteristics, max pooling is used. For classifying drowsiness, the third section utilizes gradient boosting and transfer learning/semi-supervised algorithms [37]. They reported an average of 87.46% accuracy on the NTHU-DDD dataset. The authors in [38] proposed a multi-modal and multi-view fusion model for driver facial expression recognition (FER) using RGB, Near-Infrared (NIR), and Depth Map data to address challenges posed by variations in lighting and head poses in real-world scenarios. The model, supported by a novel dataset and attention-based mechanisms, achieves over 95% accuracy, significantly outperforming single-modality approaches and demonstrating robustness in dynamic driving environments.

A common aspect of earlier systems is their reliance on one or more predefined facial features for drowsiness detection. This limits these systems' ability to detect drowsiness in situations where specific predefined facial features are not clearly visible, for example, when a driver is wearing sunglasses or a face mask, limiting the exposure of the eyes or mouth, respectively.

This work attempts to address the aforementioned limitations of previously published systems by presenting an affordable, non-intrusive, effective, real-time driver drowsiness detection system capable of working in various lighting conditions and challenging environments and scenarios.

The contributions of this paper are as follows:

- We propose, to the best of our knowledge, the first driver drowsiness detection system based on vision transformers.
- We demonstrate the effectiveness of the proposed system on challenging conditions and complex visual tasks, presented in benchmarked datasets by comparing our results with the drowsiness classification performance of previously published state-of-the-art systems trained and evaluated on those public datasets.
- We prototype and test the proposed system on hardware, making it deployable in any vehicle.

### III. PROPOSED APPROACH

#### A. CONCEPTS OF VISION TRANSFORMERS

In this section, we provide a brief overview of transformers background and working principles. Transformers were first introduced by Vaswani et al. in 2017, and they were first used for the natural language processing (NLP) domain [39]. Before transformers, NLP problems were solved using

Recurrent Neural Networks (RNNs) architectures such as LSTMs. Using RNNs, any input sequence, such as sentences, was treated sequentially to keep the order of the sequence in place. This means that each RNN layer relies on the output of the previous layer. Transformers, however, revolutionized this approach by rethinking sequence representation and introducing the concept of self-attention.

Transformers do not take ordered sequences as input; instead, they take a set of elements. In sets, order does not matter, making transformers permutation invariant. In order to keep temporal information, transformers rely on positional encoding, which is a set of small constants that will differ depending on the position of the element in the original sequence. For example, in the sentence “I love chocolate, but I hate raisins”, the two occurrences of “I” receive different positional encodings based on their respective positions in the sequence. Self-attention relates different positions of a single sequence in order to compute a representation [39]. It finds correlations between different elements in the input to associate similar elements with a higher weight.

The self-attention mechanism operates as follows: given an input matrix  $X$  and three different weight matrices  $W_q$ ,  $W_k$ , and  $W_v$ , representing query, key, and value, equations (1), (2), and (3) [39] are used to calculate  $Q$ ,  $K$ , and  $V$ , as illustrated in Figure 1. Subsequently, equation (4) [39] is then applied to calculate the attention scores given  $Q$ ,  $K$ , and  $V$ , where the ‘Softmax’ function refers to the activation function used in the output layer of the neural network model to predict the multinomial probability distribution.

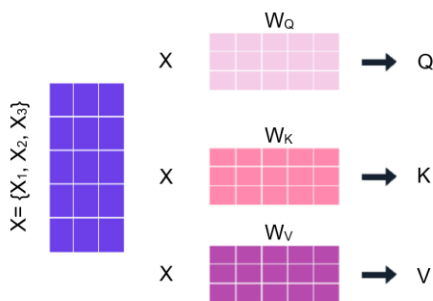
$$Q = XW_q \quad (1)$$

$$K = XW_k \quad (2)$$

$$V = XW_v \quad (3)$$

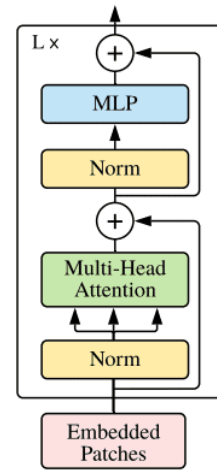
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

The attention mechanism measures the similarity between a search query and an entry key in the input sequence. Figure 2 illustrates the various components of the self-attention block. The transformer encoder consists of two layers: a multi-head attention (MHA) layer and a Multilayer



**FIGURE 1.** The input matrix  $X$  and weight matrices  $W_q$ ,  $W_k$ , and  $W_v$ , which when multiplied, result in Query ( $Q$ ), Key ( $K$ ), and Value ( $V$ ), as explained by equations (1), (2), and (3).

Perceptron (MLP) feedforward layer. The self-attention component computes attention weights for each query vector, which typically corresponds to an image patch. The MHA extends this mechanism by processing different parts of the input sequence in parallel. This attention process enables the model to extract the most important regions and distinct visual features of the image. The outputs from the independent attention heads are then linearly transformed by the MLP feedforward layer and subsequently passed to the classifier.



**FIGURE 2.** The transformer encoder block [39].

Vision Transformers (ViT) is an attempt to use the transformers method in the computer vision domain. The ViT model comprises two primary components: the backbone and the head, as shown in Figure 3. The backbone performs the encoding function, transforming input images into feature vectors, while the head is responsible for prediction, mapping the encoded feature vectors to prediction scores. Since an image is not inherently a sequence, a preprocessing step is required before transformers can be employed for image classification tasks. In 2020, a team of Google researchers introduced a method to convert the spatial non-sequential signal in images into a sequence [40]. Figure 4 depicts the steps necessary to prepare an image dataset to be fed into a transformer encoder block. As shown in the figure, before a dataset can be used to train a vision transformer architecture, each image in the dataset has to go through a conversion pipeline to make it usable by the transformer block. First, each image is split into  $n$  equal-sized batches, which are then flattened, as shown in Figure 4. Subsequently, a lower-dimensional linear embedding is applied to each patch to transform the data into a lower-dimensional feature space. The idea here is to find a reduced dimensional feature space that can effectively summarize all features in the original data. This, of course, reduces the data to be stored and processed, decreases the complexity of the model, and thus speeds up the training process. The batch order is very important to maintain the information of the original images, and thus, the next step of positional encoding is needed to give a capability



for the model to track the batch order. Finally, the dataset is fed to the transformer block shown in Figure 2.

Vision-based transformers offer several advantages and have certain limitations that should be considered. One significant advantage is their ability to capture spatial information and long-term dependencies in images, which is beneficial for image recognition and classification tasks. They also support variable-size inputs, thus making them suitable for a wide range of picture sizes. Another benefit is that they can get end-to-end training, which is more effective and enables them to learn features particular to the task. They may also be pre-trained on very large datasets to improve their performance in other tasks. However on the other hand, there are certain limitations to consider. ViT can be computationally costly, needing substantial computing power and memory. They may also suffer from fine-grained picture recognition tasks that involve catching small features. Furthermore, they often require a large amount of training data to perform effectively, which can be a practical challenge in certain applications.

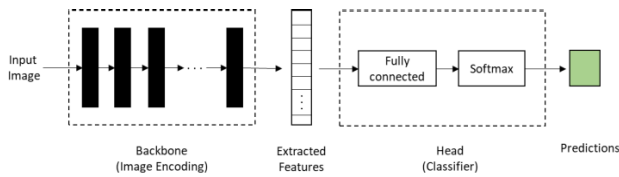


FIGURE 3. The ViT model components.

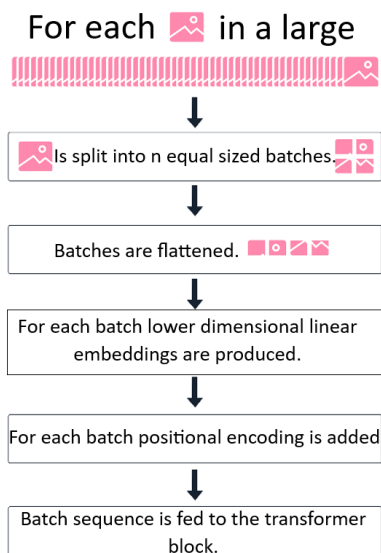


FIGURE 4. The steps taken to convert the spatial non-sequential signal in a large image dataset to a sequence usable by the transformer block.

## B. SYSTEM COMPONENTS

Our ViT-DDD system consists of two main pipelines: training and deployment. Algorithm 1 outlines the steps we adopted to prepare each of the datasets for training and evaluating our proposed model.

### Algorithm 1 Train the ViT-DDD System

---

**GIVEN (DATASET, MODEL)**  
**for** each image in **DATASET**  
     detect and crop face in image  
     save face in local desk  
  
**for** each face in local desk  
     divide face into 16x16 equal patches  
     flatten patches  
     training data.append (patches)  
  
**TRAIN (MODEL, training data)**  
  
**return** trained model

---

As detailed in Algorithm 1, before training the ViT model, we preprocess each dataset we use. The preprocessing phase involves utilizing the *dlib* face detection library [41] to detect the face of the subject in each frame of each video in the given dataset. *Dlib* is a widely used open-source, cross-platform toolkit that offers a variety of machine learning algorithms [41]. Its face detector uses the Histogram of Oriented Gradients (HOG) feature combined with a linear classifier and a sliding window detection scheme to detect human faces in images. Once the subject's face is detected, we crop it and resize the image to  $100 \times 100$  pixels, as demonstrated in Figure 5. The cropping step ensures that the ViT model only takes pixels with useful information for the drowsiness detection problem.

Since our system focuses on facial features, the region of interest (ROI) is the subject's face. Thus, we remove any background pixels, leaving only the relevant cropped face. The processed images are stored locally to eliminate the need for repetitive preprocessing during subsequent training runs on the same dataset. After preprocessing, the training data is prepared for the transformer model. Each face image from the dataset is divided into 256 equally-sized patches, as shown in Figure 6.

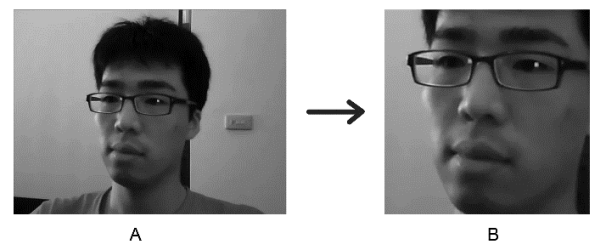
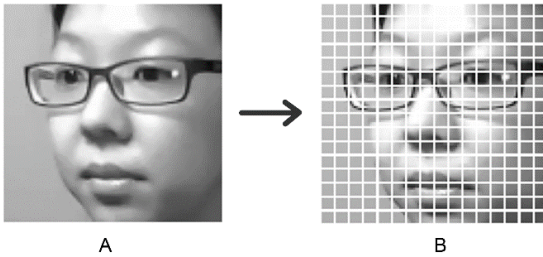


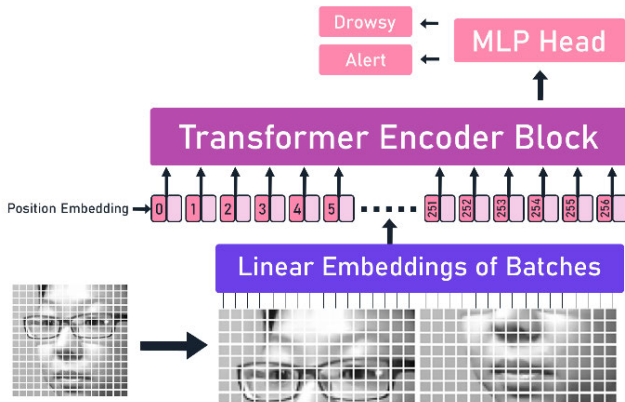
FIGURE 5. The result of the preprocessing step where (A) shows the frame of a subject of the NTHU-DDD dataset and (B) shows the detected and cropped face image saved to the local desk.

As illustrated in Figure 7, these sequences of patches are represented in a flattened vector, which is then fed into a linear projection layer that will produce a patch embedding. The purpose of the linear projection is to transform the array

of patches into lower-dimensional vectors. Each patch vector is then labeled for classification purposes. To maintain the spatial structure of the original image, positional embeddings are added to the patch embeddings. This sequence of labeled vectors is then fed into the transformer encoder. Finally, the ViT model is trained using the labeled sequences of patches from the cropped face images for each dataset, as depicted in Figure 7. The detailed structure of the proposed ViT-DDD model is shown in Figure 8. As detailed in the figure, the model is divided into three phases: (i) dividing the face image into  $n$  equal patches and linear encoding of resulting patches, (ii) transformer encoder layers, and (iii) MLP classifier. The model has 8 consecutive transformer layer blocks used for feature extraction. The last layer of the model (output layer) includes 2 neurons, with *Softmax* activation function, that predicts the state of the driver, either alert or drowsy. In addition, the hyperparameters of the model are presented in Table 1.



**FIGURE 6.** The result of dividing the face image (A) into 256 equal patches (B).



**FIGURE 7.** The ViT model training process.

During the deployment phase, as shown in Figure 9, the online monitoring pipeline applies the trained ViT model to a real-time video feed of the driver. For each frame in the video, the system detects and crops the driver's face in real time and feeds it into the trained ViT model for classification. If the system classifies a majority of frames as 'drowsy' within a given second, an alarm will be triggered to alert the driver of their condition.

**TABLE 1.** Hyperparameters of the ViT-DDD model.

Metric	Value
Input image shape	(100,100)
Num. of classes	2
Num. of patches	256
Num. of transformer layers	8
Num. of heads for attention layers	4
MLP head units	[2048, 1024]
MLP units activation function	<i>gelu</i>

#### IV. DATASETS FOR DROWSINESS DETECTION

In order to effectively train any deep learning model, a sufficient amount of relevant data is required. For drowsiness detection, we identified two suitable public datasets that have been previously used in training and evaluating driver drowsiness detection systems. This section provides a brief overview of these datasets as well as how we used them to train and evaluate our ViT-DDD model.

##### A. NTHU-DDD DATASET

The NTHU-DDD dataset [35] consists of video recordings of 36 participants, representing diverse ethnicities and genders, captured both with and without glasses or sunglasses under varying lighting conditions, including day and night. The videos were recorded using an active infrared (IR) illumination camera with a resolution of  $640 \times 480$  pixels. Participants were seated and engaged in a simulated driving game while being instructed to display a range of facial expressions such as talking, yawning, laughing, nodding, and slow blinking. These behaviors can fall into two categories: drowsy and alert. For example, yawning, nodding, slow blinking, and falling asleep are taken as drowsy. While stillness, talking, laughing, and looking aside are taken as alert. Each subject in the dataset includes five distinct scenarios to ensure variability and robustness in the analysis: no glasses, glasses, night without glasses, night with glasses, and sunglasses. In each scenario, the subject is asked to perform expressions that fall within either drowsy or normal states. Overall, 360 videos were taken to complete the dataset. After that, the dataset is organized into three distinct sections: training, evaluation, and testing. The training portion contains 18 subject folders. The evaluation portion contains 4 subject folders. We used the training and evaluation parts of the dataset to optimize the hyper-parameters of the ViT model during training by tuning them based on validation accuracy. Then, we tested our fully trained model on the testing part of the dataset. Figure 10 shows a selection of the subjects in the dataset.

##### B. UTA-RLDD DATASET

The UTA-RLDD dataset [34] comprises approximately 30 hours of RGB videos from 60 healthy participants. For each participant, the dataset includes one video corresponding to three distinct classes: "alertness", "low vigilance", and "drowsiness", totaling 180 videos. All participants were

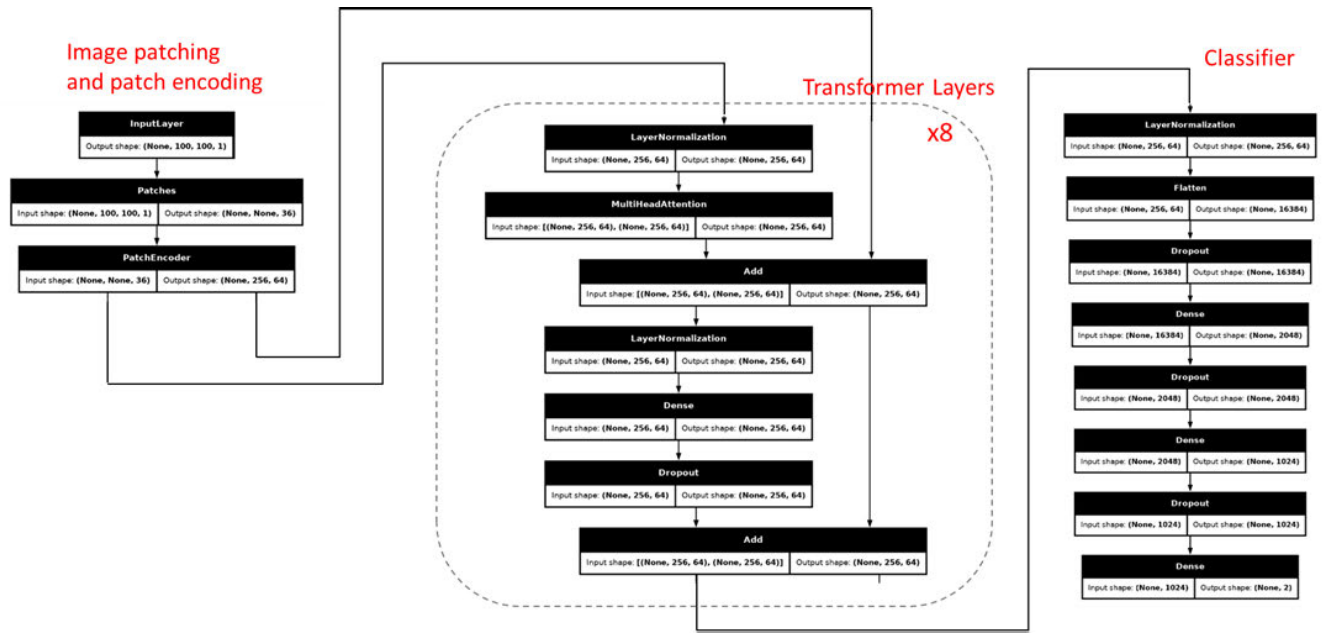


FIGURE 8. Detailed structure of the proposed ViT-DDD model.

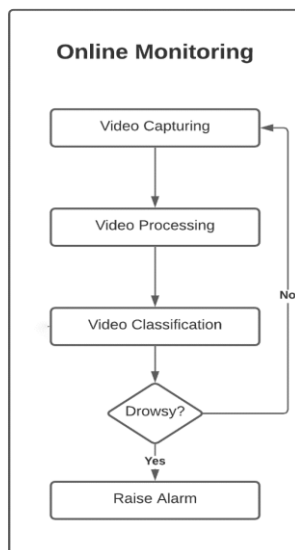


FIGURE 9. Flowchart describing the proposed alarm system.

over the age of 18, with 51 men and 9 women representing a variety of ethnic backgrounds and age groups. In 21 of the 180 videos, participants were wearing glasses, while facial hair was present in 72 videos. The videos were recorded from multiple angles, capturing diverse real-life environments and backgrounds. Importantly, all videos were self-recorded by the participants using either their smartphones or web cameras. As we aim to classify driver drowsiness for the purpose of alerting driving ideally before “drowsiness” happens, we decided to treat “low vigilance” states as part of the “drowsiness” class. This was primarily guided by the purpose of our study, which is to classify driver drowsiness for

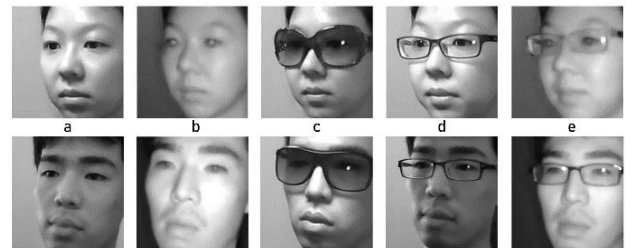
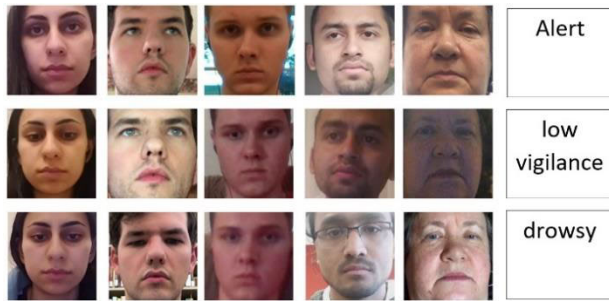


FIGURE 10. Random selection of subjects in the NTHU-DDD dataset where (a) shows subjects with no glasses in morning lighting, (b) shows subjects with no glasses at night time, (c) shows subjects with sunglasses, (d) shows subjects with glasses in morning lighting, and (e) shows subjects with glasses at night time.

safety interventions, ideally before full drowsiness happens. In the UTA-RLDD dataset, videos labeled as “low vigilance” include behavioral indicators such as some signs of sleepiness or sleepiness is present but no effort to keep alert is required. While these may not exclusively represent full drowsiness, we considered the videos of this class to fall under the drowsiness class. This will aid the developed model in detecting facial patterns and signs of fatigue at earlier stages before it becomes obvious. The UTA-RLDD dataset consists of five folds; we repetitively trained on four of the five folds and tested on the remaining one, then averaged the classification results across the five experiments. Figure 11 shows a sample of the subjects in the dataset. Table 2 presents a comparison of the key details of the NTHU-DDD and UTA-RLDD datasets.

## V. RESULTS

This section presents the experimental results of the proposed ViT-DDD system on the NTHU-DDD and



**FIGURE 11.** Random selection of different subjects in the UTA-RLDD dataset where the first row shows the subjects in their 'alert' status, the second row shows them in their 'low vigilance' state, and the third row shows them 'drowsy'.

**TABLE 2.** Comparison highlighting the key details of the NTHU-DDD and UTA-RLDD datasets.

Feature	NTHU-DDD	UTA-RLDD
Number of Participants	36 participants	60 participants
Genders of participants	50% of females	51 males and 9 females
Age of participants	From 18 to 40 years	From 20 to 59 years old, with a mean of 25 and a standard deviation of 6
Recorded Scenarios	no glasses, glasses, night without glasses, night with glasses, and sunglasses	Different signs of drowsiness
Annotations	Alert and drowsy	Alert, low vigilant, and drowsy
Dataset Size	360 videos	180 videos

UTA-RLDD datasets, concluding with a comparative analysis against previously published systems evaluated on the same datasets.

Figure 12 depicts the confusion matrix for a classifier, showing different metrics that can be used to evaluate the performance of the classifier. In our scenario, the negative class corresponds to the alert state, while the positive class corresponds to the drowsy state. The TP (True Positive) is the percentage of a driver being predicted by the classifier to be drowsy and he is actually drowsy. TN (True Negative) is the percentage of a driver being predicted to be alert and he is actually alert. FP (False Positive) corresponds to the state when the classifier predicts the driver to be drowsy while he is actually alert. Finally, FN (False Negative) corresponds to the state when the classifier predicts the driver to be alert while he is actually drowsy.

Since the NTHU-DDD dataset encompasses five distinct scenarios, we trained and evaluated our model on each scenario individually to gain a more comprehensive understanding of its performance across varying environments and lighting conditions. We trained our model on each of the five scenarios using the training section of the NTHU-DDD

dataset and tested it on each of the five scenarios in the testing section. The ViT-DDD model achieved an overall accuracy of 98.89. The confusion matrix of the classification results for each scenario is shown in Figure 13. In addition, Table 3 shows the evaluation metrics values for the proposed ViT-DDD model on the five scenarios of the NTHU-DDD dataset.

The varying lighting conditions is a challenging problem in most computer vision-based classification tasks. Most CNN models, especially those working on large objects in a scene, use lighting modification techniques, such as histogram equalization, to enhance the brightness/darkness of images which may affect the detection of these objects. However, in our work, ViTs divide an image into patches, which are treated as independent tokens. Each patch is normalized and embedded into a feature space. This patch-based representation reduces the impact of lighting variations across an image, as each patch is processed independently. Therefore, lighting variations that affect only specific regions of an image are less likely to dominate the entire feature representation. Furthermore, ViTs use a self-attention mechanism that captures global context across the image, unlike CNNs, which rely heavily on localized filters. Various studies have reported that even under small occlusions and dark lighting conditions, ViT is able to make the correct predictions [42], [43], [44].

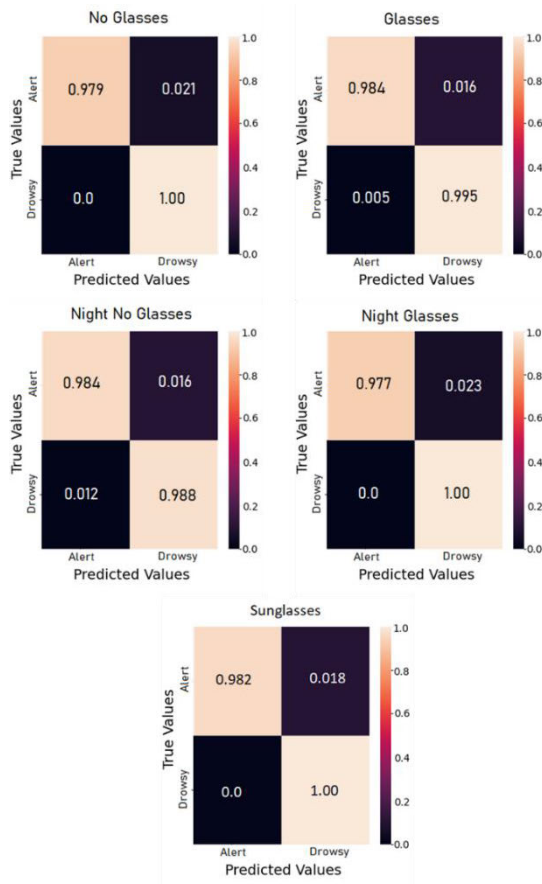
Based on the results, our ViT-DDD model demonstrated consistently high classification accuracy across all scenarios of the NTHU-DDD dataset. We believe that this robust performance can be attributed to the fact that our model does not rely on any specific facial features, such as the eyes. Instead, it builds a complete picture of the driver's status using multiple facial features, unlike the previously published systems. The fact that our model can detect drowsiness with high accuracy even when glasses and sunglasses are present and prevent a clear view of the eye suggests that the model can succeed in the challenging nature of real-life situations, which we propose as a special feature of our approach.

In Table 4, we compare the results of the proposed ViT-DDD system with three state-of-the-art systems that were trained and tested on the NTHU-DDD dataset: the HTDBN model [35], the 3D DNN model [37], and the CNN-based drowsiness detection system introduced in [45]. The later model presents our initial attempt at developing a driver drowsiness detection system in which it utilized transfer learning based on the MobileNetV2 architecture. Our previous MobileNetV2-based system starts with an image of the driver's face, which is then analyzed using the MobileNetV2 CNN architecture to extract high-level features. The resulting features are then processed by a series of MaxPooling, Flatten, Dense, and Dropout layers. The network final output identifies whether the driver is 'drowsy' or 'alert'. These models were selected for comparison with our ViT-DDD system because they use the same datasets utilized in this work and their effectiveness in driver drowsiness detection.



		Predicted Class		
		Alert (Negative)	Drowsy (Positive)	
Actual Class	Alert (Negative)	True Negative (TN)	False Positive (FP)	Specificity $\frac{TN}{TN + FP}$
	Drowsy (Positive)	False Negative (FN)	True Positive (TP)	Sensitivity $\frac{TP}{TP + FN}$
		Accuracy $\frac{TN + TP}{TN + TP + FN + FP}$	Precision $\frac{TP}{TP + FP}$	

**FIGURE 12.** The confusion matrix of a classifier, showing different performance metrics and how they are calculated.



**FIGURE 13.** The normalized confusion matrix of the ViT-DDD system on each of the five scenarios in the NTHU-DDD dataset.

As shown in Table 4, our system achieved an overall accuracy of 98.89%, outperforming other models, which recorded accuracies of 84.82%, 87.46%, and 94.39%, respectively. Notably, while the HTDBN and 3D DNN models underperformed in challenging conditions, such as the sunglasses and nightglasses scenarios, both of our models maintained high accuracies across all scenarios. This demonstrates that the adoption of the entire face, rather than focusing only

**TABLE 3.** Performance metrics for the proposed ViT-DDD on the five scenarios of the NTHU-DDD dataset.

Scenario	Specificity	Sensitivity	Precision	Accuracy
No Glasses	0.979	1.0	0.979	0.989
Glasses	0.984	0.995	0.984	0.989
Night No Glasses	0.984	0.988	0.984	0.986
Night Glasses	0.977	1.0	0.977	0.988
Sunglasses	0.982	1.0	0.982	0.991

on predefined features like the eyes or mouth, enables the detection of complex facial patterns, which contributes to the efficient detection of drowsiness. Our ViT-DDD model outperforms our earlier CNN-based system by an average of 4%.

**TABLE 4.** Comparison of classification accuracy results (%) between the HTDBN method, the 3D DNN method, our previous CNN system, and the new ViT-DDD model proposed in this paper on the five scenarios of the NTHU-DDD dataset.

Scenario	No Glasses	Glasses	Night No Glasses	Night Glasses	Sunglasses	Overall
HTDBN	92.42	86.79	91.87	75.90	76.58	84.82
3D DNN	89.12	91.36	88.33	84.39	84.16	87.46
CNN	95.82	93.74	94.63	92.56	95.21	94.39
ViT-DDD	98.97	98.93	98.61	98.84	99.12	98.89

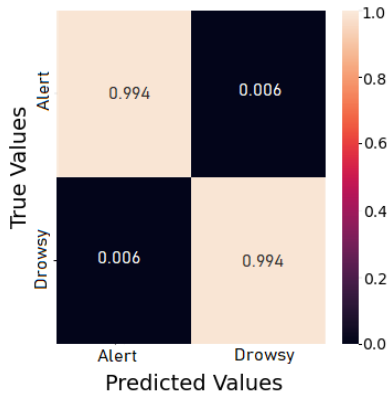
Figure 14 shows the confusion matrix of the proposed ViT-DDD system on the UTA-RLDD dataset. Moreover, the evaluation metrics results of the model are shown in Table 5.

Table 6 presents a comparison between the classification accuracy of the proposed ViT-DDD model, the HM-LSTM network introduced by the UTA-RLDD dataset creators, human experts [34], and a deep CNN model introduced by Farahnakian et al. in [36].

As shown in Table 6, our ViT-DDD system outperformed both the HM-LSTM model and human experts by a large margin of over 30%. Although our previous deep CNN approach outperformed the HM-LSTM approach and the human judgment, it fell short of our ViT-DDD system. This is mostly due to the fact that the Deep CNN method relies on the extraction of the left and right eye of the drivers along with their mouths. In contrast, the current model uses the entire facial region, resulting in a 10% improvement in accuracy, largely due to the superior feature representation of vision transformers.

These results underscore the advantage of our approach, which avoids limiting the system to specific facial features, instead allowing the model to learn subtle drowsiness cues from the entire face. This versatility is evident in our system's ability to maintain high accuracy in challenging scenarios. This is illustrated best in the fact that while other systems, such as the HTDBN and 3D DNN, struggled with

scenarios that include sunglasses or glasses at night time, our model provides consistent high-classification accuracy results in all conditions, making it more reliable in real-world usage.



**FIGURE 14.** The normalized confusion matrix of the ViT-DDD system on the UTA-RLDD dataset.

**TABLE 5.** Performance metrics for the proposed ViT-DDD model on the UTA-RLDD dataset.

Metric	Value
Specificity	0.994
Sensitivity	0.994
Precision	0.994
Accuracy	0.994

**TABLE 6.** Results comparison between human experts' judgment, the HM-LSTM network, the deep CNN method, and the ViT-DDD model proposed in this paper on the UTA-RLDD dataset.

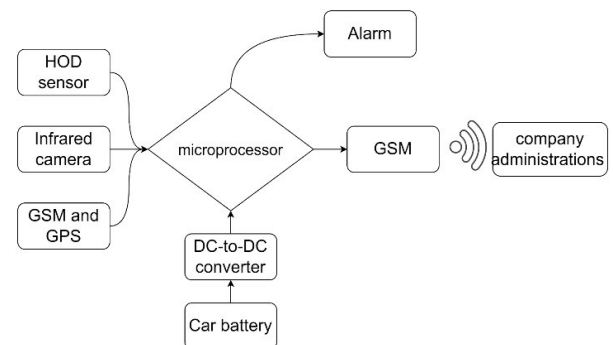
Model	Accuracy (%)
Human Judgment [34]	57.8
HM-LSTM [34]	65.2
Deep CNN [36]	89.7
<b>ViT-DDD</b>	<b>99.4</b>

This research demonstrates the effectiveness of ViT architectures in driver drowsiness detection applications. Results indicate that the proposed ViT-DDD model outperformed state-of-the-art deep learning methods, highlighting the potential of ViT in this domain. One of the key strengths of our ViT-DDD system is its ability to handle various scenarios, such as glasses, sunglasses, and different lighting conditions, by utilizing the entire facial region. The proposed system succeeded in handling these scenarios with high accuracy. Therefore, our system offers a robust and reliable solution that can be adapted to various real-world applications.

## VI. HARDWARE DESIGN AND IMPLEMENTATION

The proposed ViT-DDD system has been implemented in hardware to ensure its reliability in real scenarios. Figure 15 shows the block diagram for the system hardware design. The system comprises several components, including a Raspberry Pi 4B board [46], an infrared (IR) camera, a Hands-off Detection (HOD) sensor, a GSM/GPS module, a buck converter, and an alarm. The IR camera captures gray-scale images, which enables the system to monitor the driver's face even with poor lighting conditions or during the night. For optimal performance, the camera is fixed in front of the driver's face. The IM477-IR-CUT camera [47] was utilized in our prototype, which has a 12-megapixel resolution with a  $14 \times 18.67$ mm sensor size,  $4056 \times 3040$  active pixels, and 4k/60fps video recording capabilities. It operates with a 3.3V power supply and is equipped with an IR-cut filter for improved image quality. The Raspberry Pi 4B microcomputer offers robust processing and storage capabilities with a 64-bit processor, 1.5 GHz CPU, and 8 GB RAM, making it ideal for this video processing task.

In addition, an FSR-408 sensor [48] is used in our prototype as a HOD sensor to monitor the driver's hands on the steering wheel. The FSR-408 sensor is a resistive-based sensor that is easy to install on the car steering wheel and works even when the driver is wearing gloves, as it depends on the amount of pressure on the wheel. The HOD sensor measures the force applied to the steering wheel to determine whether the driver's hands are on or off the steering wheel. This provides additional valuable information regarding the driver's drowsiness state. As shown in Figure 15, the camera, the HOD sensor, and the GSM/GPS module are all connected to the Raspberry Pi, which is powered by the car's battery. A DC-to-DC converter is used to reduce the 12V car battery voltage to the microprocessor's operating voltage. A buzzer is also used to alert the driver; however, if the driver does not respond, a message would be sent through the GSM/GPS module to the vehicle owner (e.g., the company in case of a taxi or bus driver).



**FIGURE 15.** Block diagram of the proposed system implementation.

First, the model is trained on a computer, and the trained model, along with the optimum weights, is compiled for deployment on the Raspberry Pi board for real-life

monitoring. In the second phase, the camera continuously monitors the driver's face and supplies the captured images to the model for classification to determine whether the driver is exhibiting signs of drowsiness, as demonstrated in Figure 16. Additionally, the HOD sensor functions concurrently with the camera. If the system detects that the driver's hand is off the steering wheel or that the driver appears drowsy, an alert is triggered. If the driver fails to respond to the alert, the system sends a notification to corporate management via the GPS/GSM module, including the driver's location and a warning message.

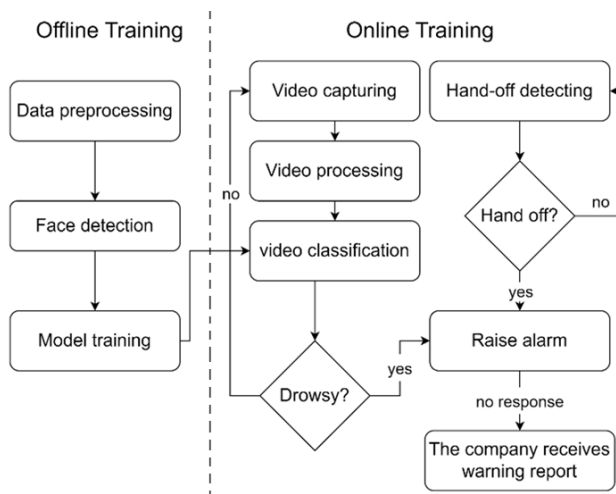


FIGURE 16. Flowchart of the system software.

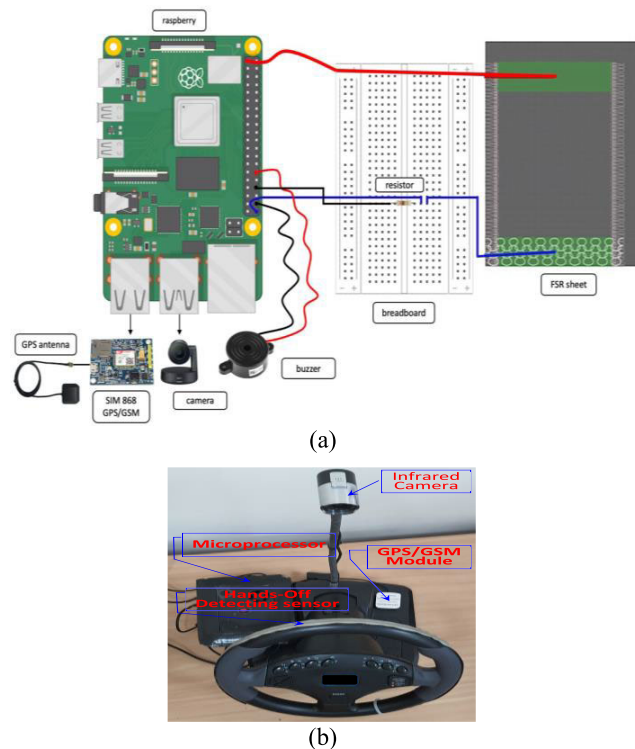


FIGURE 17. (a) Implemented system; and (b) system prototype.

Execution time is a critical issue for this system, which should be fast enough to detect the driver's drowsiness and take the necessary action to avoid an accident. This issue has been addressed in this work by reducing the number of captured frames to 4 frames per second, which is enough to monitor the driver's face and differentiate between 'normal' and 'drowsy' driving. Using more than 4 frames per second would overload the microprocessor and cause it to lag.

Figure 17 shows the hardware prototype for the whole system developed as part of this work to test the performance of the system in a practical environment. A real scenario to test the performance of the system in a practical driving environment is shown in Figure 18, in which the system succeeded to differentiate between normal and drowsy states. In addition, if the drowsiness state was detected and no action was taken by the driver, an alert message will be sent to the corporate management, including the driver's location, as shown in Figure 19.



FIGURE 18. Testing the hardware system in a practical environment, which is able to differentiate between normal and drowsy driving.

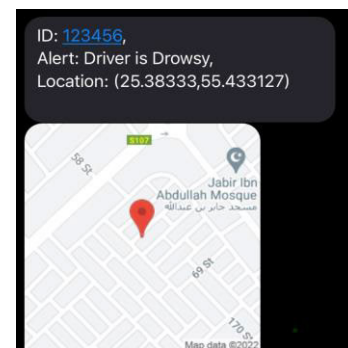


FIGURE 19. A warning message was sent to corporate management indicating that the driver is drowsy, including the driver's location.

## VII. CONCLUSION AND FUTURE WORK

In this paper, we presented a real-time and non-intrusive Vision Transformer-based Driver Drowsiness Detection (ViT-DDD) system designed to address the critical issue of detecting drowsiness in real time to prevent accidents caused by drowsy and fatigued drivers. The system was tested on two public datasets: NTHU-DDD and UTA-RLDD. Results showed that the proposed system achieves superior performance compared to traditional CNN models, particularly in

challenging conditions such as low-light environments or when the driver is wearing sunglasses. The utilization of entire face images, rather than relying on isolated features, such as the eyes or mouth, enabled our model to detect subtle indicators of drowsiness with high accuracy across different scenarios. The model showed an increase with more than 4% in accuracy compared to other related studies in the literature. In addition, the proposed system was implemented on a Raspberry Pi microcomputer to assess its performance in real scenarios. The hardware system includes an IR camera, HOD sensor, and GPS/GSM module. The prototype system effectively monitors the driver's facial expressions and hand position on the steering wheel in real-time, issuing timely alerts and escalating notifications if necessary, demonstrating the system's practical application.

Some limitations of the current work include the possibility that our system may fail if a significant portion of the face is obscured. Also, the transmission of the alert message may fail if the area has no GPS or GSM coverage, which affects the timely response of the system. While the performance of the system may deteriorate if a majority of the face is obscured or the face is out of the frame during processing, we may incorporate in the future an additional warning message to the driver asking him to remove any obstructions on his face that may affect the working of the system or adjust his position when a face is not detected.

At this stage, our system has been tested in lab conditions using publicly available datasets. Future avenues for improving our ViT-DDD system include utilizing temporal information and the spatial information we already use from each frame. Using temporal information should result in a more comprehensive drowsiness detection system that is capable of relating the subtle clues of a driver's face over time, which should result in higher classification performance and the ability to predict drowsiness in advance. In addition, the experiments were designed using custom ViT configurations (8 transformer layers and 256 patches for  $100 \times 100$  pixels images). However, we acknowledge the importance of systematically analyzing how varying these parameters affects the model. While we did not explicitly vary these parameters in our experiments, we plan to conduct a comparison of these parameters in our future work for other applications. Future work would also include systematic testing of the system based on the implemented prototype in order to establish its effectiveness and applicability in practical situations. Additionally, future work will explore integrating this system with other relevant and available vehicle metrics to provide a holistic safety solution, capable of reacting not only to signs of drowsiness but also to other risky driving behaviors.

## DATA AVAILABILITY

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

## DECLARATIONS

### CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

## REFERENCES

- [1] *Summary of Hours of Service Regulations* | FMCSA. Accessed: Dec. 5, 2022. [Online]. Available: <https://www.fmcsa.dot.gov/regulations/hours-service/summary-hours-service-regulations>
- [2] *Are You Sleeping Enough? This Infographic Shows How You Compare To the Rest of the World*. World Economic Forum. Accessed: Dec. 5, 2022. [Online]. Available: <https://www.weforum.org/agenda/2019/08/we-need-more-sleep/>
- [3] *Drowsy Driving-19 States and the District of Columbia, 2009–2010*. Accessed: Dec. 5, 2022. [Online]. Available: <https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6151a1.htm>
- [4] A. G. Wheaton, R. A. Shults, D. P. Chapman, E. S. Ford, and J. B. Croft, "Drowsy driving and risk behaviors—10 states and Puerto Rico, 2011–2012," *MMWR Morb. Mortal. Wkly. Rep.*, vol. 63, no. 26, pp. 557–562, Jul. 2014.
- [5] *Drowsy Driving* | NHTSA. Accessed: Dec. 8, 2022. [Online]. Available: <https://www.nhtsa.gov/risky-driving/drowsy-driving>
- [6] (2006). *Institute of Medicine (US) Committee on Sleep Medicine and Research, Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem*. in *The National Academies Collection: Reports Funded By National Institutes of Health*. National Academies Press, Washington, DC, USA. Accessed: Jun. 5, 2023. [Online]. Available: <http://www.ncbi.nlm.nih.gov/books/NBK19960/>
- [7] B. C. Tefft, "Prevalence of motor vehicle crashes involving drowsy drivers, United States, 1999–2008," *Accident Anal. Prevention*, vol. 45, pp. 180–186, Mar. 2012, doi: [10.1016/j.aap.2011.05.028](https://doi.org/10.1016/j.aap.2011.05.028).
- [8] A. I. Siam, S. A. Gamel, and F. M. Talaat, "Automatic stress detection in car drivers based on non-invasive physiological signals using machine learning techniques," *Neural Comput. Appl.*, vol. 35, no. 17, pp. 12891–12904, Jun. 2023, doi: [10.1007/s00521-023-08428-w](https://doi.org/10.1007/s00521-023-08428-w).
- [9] S. A. El-Nabi, W. El-Shafai, E.-S.-M. El-Rabaie, K. F. Ramadan, F. E. A. El-Samie, and S. Mohsen, "Machine learning and deep learning techniques for driver fatigue and drowsiness detection: A review," *Multimedia Tools Appl.*, vol. 83, no. 3, pp. 9441–9477, Jan. 2024, doi: [10.1007/s11042-023-15054-0](https://doi.org/10.1007/s11042-023-15054-0).
- [10] K. Fujiwara, E. Abe, K. Kamata, C. Nakayama, Y. Suzuki, T. Yamakawa, T. Hiraoka, M. Kano, Y. Sumi, F. Masuda, M. Matsuo, and H. Kadotani, "Heart rate variability-based driver drowsiness detection and its validation with EEG," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 6, pp. 1769–1778, Jun. 2019, doi: [10.1109/TBME.2018.2879346](https://doi.org/10.1109/TBME.2018.2879346).
- [11] V. P. Balam, "Systematic review of single-channel EEG-based drowsiness detection methods," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 11, pp. 15210–15228, Nov. 2024, doi: [10.1109/TITS.2024.3442249](https://doi.org/10.1109/TITS.2024.3442249).
- [12] R. Alharbey, M. M. Dessouky, A. Sedik, A. I. Siam, and M. A. Elaskily, "Fatigue state detection for tired persons in presence of driving periods," *IEEE Access*, vol. 10, pp. 79403–79418, 2022, doi: [10.1109/ACCESS.2022.3185251](https://doi.org/10.1109/ACCESS.2022.3185251).
- [13] U. Budak, V. Bajaj, Y. Akbulut, O. Atila, and A. Sengur, "An effective hybrid model for EEG-based drowsiness detection," *IEEE Sensors J.*, vol. 19, no. 17, pp. 7624–7631, Sep. 2019, doi: [10.1109/JSEN.2019.2917850](https://doi.org/10.1109/JSEN.2019.2917850).
- [14] A. Kolus, "A systematic review on driver drowsiness detection using eye activity measures," *IEEE Access*, vol. 12, pp. 97969–97993, 2024, doi: [10.1109/ACCESS.2024.3424654](https://doi.org/10.1109/ACCESS.2024.3424654).
- [15] K. Fujiwara, H. Iwamoto, K. Hori, and M. Kano, "Driver drowsiness detection using R-R interval of electrocardiogram and self-attention autoencoder," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 1, pp. 2956–2965, Jan. 2024, doi: [10.1109/TIV.2023.3308575](https://doi.org/10.1109/TIV.2023.3308575).
- [16] M. Awais, N. Badruddin, and M. Driberg, "A hybrid approach to detect driver drowsiness utilizing physiological signals to improve system performance and wearability," *Sensors*, vol. 17, no. 9, p. 1991, Aug. 2017, doi: [10.3390/s17091991](https://doi.org/10.3390/s17091991).
- [17] K. Singh. (2013). *Physical and Physiological Drowsiness Detection Methods*. Accessed: Jun. 5, 2023. [Online]. Available: <https://www.semanticscholar.org/paper/Physical-and-Physiological-Drowsiness-Detection-Singh/d7d65e7e72c7811708ab578dbd83c7c278892d5c>



- [18] *Driver Alert System*. Volkswagen Newsroom. Accessed: Dec. 10, 2022. [Online]. Available: <https://www.volkswagen-newsroom.com/en/driver-alert-system-3932>
- [19] *How Does Mercedes Benz Attention Assist Work? | Technology Overview*. Mercedes-Benz of Easton. Accessed: Jun. 5, 2023. [Online]. Available: <https://www.mercedesbenzofeaston.com/mercedes-benz-attention-assist/>
- [20] J. Vicente, P. Laguna, A. Bartra, and R. Bailón, "Drowsiness detection using heart rate variability," *Med. Biol. Eng. Comput.*, vol. 54, no. 6, pp. 927–937, Jun. 2016, doi: [10.1007/s11517-015-1448-7](https://doi.org/10.1007/s11517-015-1448-7).
- [21] E. Khosravi, A. M. A. Hemmatyar, M. J. Siavoshani, and B. Moshiri, "Safe deep driving behavior detection (S3D)," *IEEE Access*, vol. 10, pp. 113827–113838, 2022, doi: [10.1109/ACCESS.2022.3217644](https://doi.org/10.1109/ACCESS.2022.3217644).
- [22] G. Sai Krishna, K. Supriya, J. Vardhan, and M. Rao, "Vision transformers and YoloV5 based driver drowsiness detection framework," 2022, *arXiv:2209.01401*.
- [23] K. Peng, A. Roitberg, K. Yang, J. Zhang, and R. Stiefelhagen, "TransDARC: Transformer-based driver activity recognition with latent space feature calibration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2022, pp. 278–285, doi: [10.1109/iros47612.2022.9981445](https://doi.org/10.1109/iros47612.2022.9981445).
- [24] C. B. S. Maior, M. J. D. C. Moura, J. M. M. Santana, and I. D. Lins, "Real-time classification for autonomous drowsiness detection using eye aspect ratio," *Exp. Syst. Appl.*, vol. 158, Nov. 2020, Art. no. 113505, doi: [10.1016/j.eswa.2020.113505](https://doi.org/10.1016/j.eswa.2020.113505).
- [25] M. Ngxande, J.-R. Tapamo, and M. Burke, "Driver drowsiness detection using behavioral measures and machine learning techniques: A review of state-of-art techniques," in *Proc. Pattern Recognit. Assoc. South Afr. Robot. Mechatronics*, Nov. 2017, pp. 156–161, doi: [10.1109/RoboMech.2017.8261140](https://doi.org/10.1109/RoboMech.2017.8261140).
- [26] E. Tadesse, W. Sheng, and M. Liu, "Driver drowsiness detection through HMM based dynamic modeling," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 4003–4008, doi: [10.1109/ICRA.2014.6907440](https://doi.org/10.1109/ICRA.2014.6907440).
- [27] H. Yin, Y. Su, Y. Liu, and D. Zhao, "A driver fatigue detection method based on multi-sensor signals," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–7, doi: [10.1109/WACV.2016.7477672](https://doi.org/10.1109/WACV.2016.7477672).
- [28] L. K. McIntire, R. A. McKinley, C. Goodyear, and J. P. McIntire, "Detection of vigilance performance using eye blinks," *Appl. Ergonom.*, vol. 45, no. 2, pp. 354–362, Mar. 2014, doi: [10.1016/j.apergo.2013.04.020](https://doi.org/10.1016/j.apergo.2013.04.020).
- [29] A. Narayanan, R. M. Kaimal, and K. Bijlani, "Estimation of driver head yaw angle using a generic geometric model," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 12, pp. 3446–3460, Dec. 2016, doi: [10.1109/TITS.2016.2551298](https://doi.org/10.1109/TITS.2016.2551298).
- [30] B. Akrouf and W. Mahdi, "A novel approach for driver fatigue detection based on visual characteristics analysis," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 1, pp. 527–552, Jan. 2023, doi: [10.1007/s12652-021-03311-9](https://doi.org/10.1007/s12652-021-03311-9).
- [31] M. Dua, R. Singla, S. Raj, and A. Jangra, "Deep CNN models-based ensemble approach to driver drowsiness detection," *Neural Comput. Appl.*, vol. 33, no. 8, pp. 3155–3168, Apr. 2021, doi: [10.1007/s00521-020-05209-7](https://doi.org/10.1007/s00521-020-05209-7).
- [32] R. Jabbar, K. Al-Khalifa, M. Kharbeche, W. Alhajyaseen, M. Jafari, and S. Jiang, "Real-time driver drowsiness detection for Android application using deep neural networks techniques," *Proc. Comput. Sci.*, vol. 130, pp. 400–407, Jan. 2018, doi: [10.1016/j.procs.2018.04.060](https://doi.org/10.1016/j.procs.2018.04.060).
- [33] R. Jabbar, M. Shinoy, M. Kharbeche, K. Al-Khalifa, M. Krichen, and K. Barkaoui, "Driver drowsiness detection model using convolutional neural networks techniques for Android application," in *Proc. IEEE Int. Conf. Informat., IoT, Enabling Technol. (ICIOT)*, Feb. 2020, pp. 237–242, doi: [10.1109/ICIOT48696.2020.9089484](https://doi.org/10.1109/ICIOT48696.2020.9089484).
- [34] R. Ghoddoosian, M. Galib, and V. Athitsos, "A realistic dataset and baseline temporal model for early drowsiness detection," presented at the *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019. [Online]. Available: [https://openaccess.thecvf.com/content\\_CVPRW\\_2019/html/AMFG/Ghoddoosian\\_A\\_Realistic\\_Dataset\\_and\\_Baseline\\_Temporal\\_Model\\_for\\_Early\\_Drowsiness\\_CVPRW\\_2019\\_paper.html](https://openaccess.thecvf.com/content_CVPRW_2019/html/AMFG/Ghoddoosian_A_Realistic_Dataset_and_Baseline_Temporal_Model_for_Early_Drowsiness_CVPRW_2019_paper.html)
- [35] C.-H. Weng, Y.-H. Lai, and S.-H. Lai, "Driver drowsiness detection via a hierarchical temporal deep belief network," in *Computer Vision—ACCV (Lecture Notes in Computer Science)*, C.-S. Chen, J. Lu, and K.-K. Ma, Eds., Cham, Switzerland: Springer, 2017, pp. 117–133, doi: [10.1007/978-3-319-54526-4\\_9](https://doi.org/10.1007/978-3-319-54526-4_9).
- [36] F. Farahnakian, J. Leoste, and F. Farahnakian, "Driver drowsiness detection using deep convolutional neural network," in *Proc. Int. Conf. Electr., Comput., Commun. Mechatronics Eng. (ICECCME)*, Oct. 2021, pp. 1–6, doi: [10.1109/ICECCME52200.2021.9591029](https://doi.org/10.1109/ICECCME52200.2021.9591029).
- [37] X.-P. Huynh, S.-M. Park, and Y.-G. Kim, "Detection of driver drowsiness using 3D deep neural network and semi-supervised gradient boosting machine," in *Computer Vision—ACCV (Lecture Notes in Computer Science)*, C.-S. Chen, J. Lu, and K.-K. Ma, Eds., Cham, Switzerland: Springer, 2017, pp. 134–145, doi: [10.1007/978-3-319-54526-4\\_10](https://doi.org/10.1007/978-3-319-54526-4_10).
- [38] J. Chen, S. Dey, L. Wang, N. Bi, and P. Liu, "Attention-based multi-modal multi-view fusion approach for driver facial expression recognition," *IEEE Access*, vol. 12, pp. 137203–137221, 2024, doi: [10.1109/ACCESS.2024.3462352](https://doi.org/10.1109/ACCESS.2024.3462352).
- [39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.
- [40] A. Dosovitskiy, "An image is worth 16×16 words: Transformers for image recognition at scale," Presented at the Int. Conf. Learn. Represent., Jan. 2021. [Online]. Available: <https://openreview.net/forum?id=YicbFdNTTy>
- [41] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, no. 60, pp. 1755–1758, Dec. 2009.
- [42] S. Paul and P. Chen, "Vision transformers are robust learners," in *Proc. AAAI Conf. Artif. Intell.*, Jan. 2021, pp. 2071–2081.
- [43] M. Naseer, K. Ranasinghe, S. Khan, M. Hayat, F. S. Khan, and M. Yang, "Intriguing properties of vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2021, pp. 23296–23308.
- [44] D. Zhou, Z. Yu, E. Xie, C. Xiao, A. Anandkumar, J. Feng, and J. M. Álvarez, "Understanding the robustness in vision transformers," in *Proc. 39th Int. Conf. Mach. Learn., Proc. Mach. Learn. Res.*, Jan. 2022, pp. 27378–27394. [Online]. Available: <https://proceedings.mlr.press/v162/zhou22m.html>
- [45] M. S. Mahmoud, A. Jarndal, A. Alzghoul, H. Almahasneh, I. Alsyouf, and A. K. Hamid, "Driver drowsiness detection system using deep learning based on visual facial features," in *Proc. 14th Int. Conf. Develop. eSystems Eng. (DeSE)*, Dec. 2021, pp. 453–458, doi: [10.1109/DeSE54285.2021.9719409](https://doi.org/10.1109/DeSE54285.2021.9719409).
- [46] *Raspberry Pi 4 Model B Specifications*. Raspberry PI. Accessed: Oct. 9, 2022. [Online]. Available: <https://www.raspberrypi.com/products/raspberry-pi-4-model-b/specifications/>
- [47] *ArduCAM B0270 Raspberry Pi IR-CUT HQ Camera With CS-Mount Lens User Manual*. Accessed: Jun. 1, 2023. [Online]. Available: <https://manuals.plus/arducam/b0270-raspberry-pi-ir-cut-hq-camera-with-cs-mount-lens-manual>
- [48] *FSR 408*. Accessed: Jun. 1, 2023. [Online]. Available: <https://www.interlinkelectronics.com/fsr-408>



**ANWAR JARNDAL** (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Kassel, Germany, in 2006. He was a Postdoctoral Fellow with École de Technologie Supérieure (ETS), Quebec University, Canada. He is currently a Professor with the Department of Electrical Engineering, University of Sharjah. Since 2020, he has been ranked among the top 2% of lifetime-cited researchers, according to Stanford University's analysis. He has published more than 150 internationally peer-reviewed articles and serves as a reviewer for more than 30 international journals. His research interests include active device modeling, measurement and characterization techniques, power amplifier design, low-noise amplifier design, local and global optimization, artificial neural networks, machine learning, fuzzy logic, radio channel modeling, and wireless power transfer. He has been honored with the University of Sharjah's Annual Incentive Award for Distinguished Faculty in Scientific Research.



**HISSAM TAWFIK** received the Ph.D. degree in electrical and computer engineering from The University of Manchester, U.K., and has a well-established research track record of refereed publications in reputable international journals and conference proceedings in artificial intelligence for engineering, biomedical engineering and biologically inspired systems, big data and machine learning. He is currently a Professor of artificial intelligence. He is a Professor with the College of Engineering, University of Sharjah, United Arab Emirates. Prior to that, he worked for more than 20 years for various universities in U.K. He served as an Editor of *International Journal of Future Generation Computer Systems* (Elsevier), and *International Journal of Reliable Intelligent Environments* (Springer), and is an Editor for *International Journal of Neural Computing and Applications* (Springer). He is the Chair of the International Conference Series on Developments in eSystems Engineering (DESE) and a Guest Editor of the Special Collection on “Robotics, Sensors, and Industry 4.0” for the *Journal of Sensors*, MPDI.



**ALI I. SIAM** received the Ph.D. degree in electronics and electrical communications engineering from the Faculty of Electronic Engineering, Menoufia University, Egypt, in 2021. He is currently an Assistant Professor with the Department of Embedded Network Systems Technology, Faculty of Artificial Intelligence, Kafrelsheikh University, Egypt. He is a Postdoctoral Fellow with the University of Sharjah, United Arab Emirates. His research interests include machine learning, deep learning, computer vision, system faults identification, the IoT, system security, and signal and image processing.



**IMAD ALSAYOUF** received the Ph.D. degree in industrial engineering from Linnaeus University, Sweden, in 2004. He has had a distinguished 35-year career that spans both industry and academia. He served for approximately ten years with Linnaeus University, where he was promoted to an Associate Professor before joining the University of Sharjah (UOS), in 2010. From 2015 to 2016, he was the Head of the Industrial Engineering and Engineering Management Department, UOS. He founded and led the Sustainable Engineering Asset Management (SEAM) Research Group, from 2016 to 2021. Additionally, he established and led the Sustainability Office, UOS, from 2017 to 2024, where he introduced the innovative concept of sustainability circles. Throughout his career, he has supervised numerous Ph.D. and master's students, authored more than 100 Scopus-indexed publications, and has been recognized among the top 2% of cited scientists by Stanford University, in 2022 and 2023. His contributions not only advance sustainability but also establish him as a leading expert in his field.



**ALI CHEAITOU** is currently an Associate Professor of industrial engineering and engineering management, and the Coordinator of the Sustainable Engineering Asset Management (SEAM) Research Group, University of Sharjah, United Arab Emirates. Previously, he was the Chairman of the Department of Industrial Engineering and Engineering Management, from 2018 to 2022 and as the Coordinator of the M.Sc. and Ph.D. Programs in engineering management with the University of Sharjah, from 2013 to 2017. Before joining the University of Sharjah, he was an Assistant Professor with Euromed Management (Kedge Business School), Marseilles, France, and as a Lecturer with Paris Saclay University (École Centrale Paris), France. He also spent two years in the industry as an ERP and supply chain management consultant, mainly with L'Oréal, Paris, France. His main research interests include sustainable supply chain management and optimization of logistics systems, with a focus on transportation. He is an Associate Editor of *Supply Chain Forum: An International Journal*.

...