Identification Detection and YoloV5 based Driver Drowsiness Framework

Abdul Rehman[1]; Mohammad Zaidan Waseem[2]; Abdul Rafey[3];
Ali Abbas Hussaini[4]; Hina Parveen[5]; Nida Khan[6]

Department of Computer Science Integral University, Lucknow

Publication Date: 2025/05/08

Abstract; Human drivers bring a unique mix of skills, instincts, and emotions to the road, shaped by their individual driving habits. However, driver drowsiness poses a serious threat to road safety, making it critical to develop effective detection systems to prevent accidents. Past efforts to identify unusual driver behavior often focused on analyzing the driver's face or vehicle movements using computer vision techniques. While these methods provided some insights, they struggled to capture the full complexity of driver behavior.

With the rise of deep learning, researchers have increasingly turned to neural networks to better understand and detect driver drowsiness. This paper presents a fresh approach using vision transformers and YOLOv5 architectures to recognize signs of drowsiness. We propose a customized YOLOv5 model, pre-trained to detect and extract the driver's face as the Region of Interest (ROI). To overcome the limitations of earlier systems, we incorporate vision transformers for binary image classification. Our model was trained and tested on the public UTA-RLDD dataset, achieving impressive results with 96.2% training accuracy and 97.4% validation accuracy.

To further evaluate its performance, we tested the framework on a custom dataset of 39 participants under various lighting conditions, where it achieved a solid 95.5% accuracy. These experiments highlight the strong potential of our approach for real-world use in smart transportation systems, paving the way for safer roads.

Keywords; *Drowsiness Detection, Image Classification , Vision Transformers (VIT) ,Yolov5, Face Detection.*

How to Cite: Abdul Rehman; Mohammad
Zaidan Waseem; Abdul Rafey; Ali Abbas Hussaini; Hina Parveen; Nida Khan (2025). Identification Detection and YoloV5 based Driver Drowsiness Framework. *International Journal of Innovative Science and Research Technology*, 10(4), 2806-2815. https://doi.org/10.38124/ijisrt/25apr1476

I. INTRODUCTION

The lack of effective drowsiness detection in Advanced Driver Assistance Systems (ADAS) remains a major contributor to road accidents, leading to severe harm for both drivers and pedestrians. The Central Road Research Institute (CRRI) reports that fatigue-related crashes—often caused when drivers fall asleep behind the wheel—are responsible for a significant portion of traffic fatalities and nearly 40% of road injuries. Data from the National Highway Traffic Safety Administration (NHTSA) [1] reveals that drowsy driving causes approximately one million crashes each year, resulting in around 2,000 fatalities and 70,000 injuries. Alarmingly, nearly 80% of these accidents involve single-vehicle run-off-road incidents, where the driver loses control and either leaves the road or collides with another vehicle [2]. This underscores how drowsiness, although frequently underestimated, poses a serious risk on the roads and highlights the growing need for reliable detection systems to enhance safety.

In recent times, detecting driver fatigue has become a prominent area of research. Drowsiness detection methods are generally grouped into three categories: physiological monitoring, vehicle-based analysis, and facial behavior tracking [3], [4]. Physiological techniques involve tracking body signals like heart rate, brain activity, and skin temperature, which shift as a person becomes sleepy [5]. Technologies such as electrocardiography (ECG) [6], electroencephalography (EEG) [7], electromyography (EMG), and electrooculography (EOG) [8] are commonly used to assess a driver's condition. A significant challenge in this approach, however, is maintaining user comfort while collecting data through sensors [9]. Vehicle-based strategies monitor driving behavior—such as erratic steering, sudden braking, or inconsistent speed—to identify signs of fatigue. Sensors within the car record these patterns to detect unusual activity. Still, this method can be affected by external variables like bad weather, rough roads, or medications taken by the driver [10]. Meanwhile, facial analysis provides a contactless solution by using computer vision and machine learning to observe facial movements and expressions that indicate drowsiness [11].

Recently, several behavior-based methods have been developed for detecting fatigue in drivers. Among the most widely used techniques are facial recognition models like the Viola-Jones (Haar Cascade) algorithm [12], Canny edge detection [3], and various neural networks including Convolutional Neural Networks (CNN) [13], Artificial Neural Networks (ANN) [14], Naive Bayes classifiers [12], and Generative Adversarial Networks (GANs) [15]. These models focus on analyzing visual features to identify fatigue with high accuracy. In this research, we propose a behavioral detection system that utilizes both face detection and drowsiness recognition by combining YOLOv5 and Vision Transformers (ViT). The key contributions of this work are as follows:

- ➢ *Custom training of a YOLOv5 model specifically for accurate face detection.*
- ➢ *Design and evaluation of a binary image classification system using Vision Transformers for identifying drowsiness.*
- ➢ *Real-time performance testing of the model on a specially curated dataset with diverse driving conditions.*

The remainder of this paper is structured as follows: Section II reviews prior research on drowsiness detection techniques. Section III outlines the proposed approach, beginning with face detection using YOLOv5, then moving to data augmentation, and finally, image classification with Vision Transformers. Section IV discusses the experiments and results, including model accuracy on our custom dataset. Section V presents the main findings and compares them with existing methods. Section VI concludes the paper with final thoughts and future directions.

II. RELATED WORKS

Driving plays a vital role in everyday life, which makes it essential to thoroughly understand, analyze, and anticipate driver behavior. Recognizing the importance of this, many researchers have explored techniques to identify irregular behaviors, particularly driver drowsiness. Recently, there has been a noticeable shift toward more sophisticated approaches to improve the accuracy of drowsiness detection systems. For example, Zuojin Li et al. [16] investigated vehicle-centric methods by gathering yaw angle and steering wheel data through sensors mounted in vehicles. They calculated approximate entropy from this time-series data and used these features to train a Back-propagation Neural Network. Their system classified driver states—awake, drowsy, or very drowsy—with an accuracy of 87.21%.

To overcome the limitations found in physiological and vehicle-based methods, researchers have increasingly focused on behavioral approaches. While vehicle-based methods rely on the vehicle’s movement patterns, behavioral techniques concentrate on the driver’s facial cues, which tend to yield more consistent results. Although physiological monitoring offers high accuracy, it often lacks practicality due to its complex setup. Advancing this field, Sherif Said et al. [17] designed a system using the Viola-Jones algorithm to detect facial and eye regions. Their system activated an alert when signs of drowsiness were detected, achieving an accuracy of 82% indoors and 72.8% outdoors. Likewise, Feng You [18] proposed a two-phase model that starts with offline training and transitions to real-time use. The method uses DLIB’s CNN for face detection and calculates the eye aspect ratio based on 68 facial landmarks. The offline stage involves Support Vector Machine (SVM) training, followed by online monitoring, and achieves a high accuracy of 94.8%. However, its limitation lies in the need to train the SVM individually for each user. Another study [15] enhanced performance by using Generative Adversarial Networks (GANs) alongside Convolutional Neural Networks (CNNs) to generate synthetic training data, resulting in improved model accuracy.

In a different approach, R. Tamanani et al. [19] introduced a system made up of two sequential components. The first stage uses the Haar Cascade algorithm for facial detection and real-time data processing, while the second stage utilizes a CNN model for extracting features and performing classification. Their model, tested on the UTA-RLDD dataset using 5-fold stratified cross-validation, reached impressive metrics—precision, recall, and F1 scores of 91.8%, 92.8%, and 92%, respectively. When evaluated on a custom dataset, it achieved training, validation, and testing accuracies of 98%, 84%, and 88%.

In this research, we propose a novel framework that integrates YOLOv5 [20], [21] with Vision Transformers [22], [23], [24], [25] to offer a more robust alternative to existing approaches. The system’s real-time performance was evaluated using a custom-built dataset designed to simulate various real-world driving scenarios.

III. RESEARCH METHODOLOGY

This section introduces our novel framework, which is composed of several well-defined stages. At its core, the system incorporates a fine-tuned, pretrained YOLOv5 model for automatic face detection and a custom-trained Vision Transformer (ViT) model for binary image classification. To strengthen the performance of the ViT classifier, we employed various data augmentation strategies to enlarge and diversify the training dataset. An outline of the complete framework, including the drowsiness detection components, is illustrated in Figure 1.

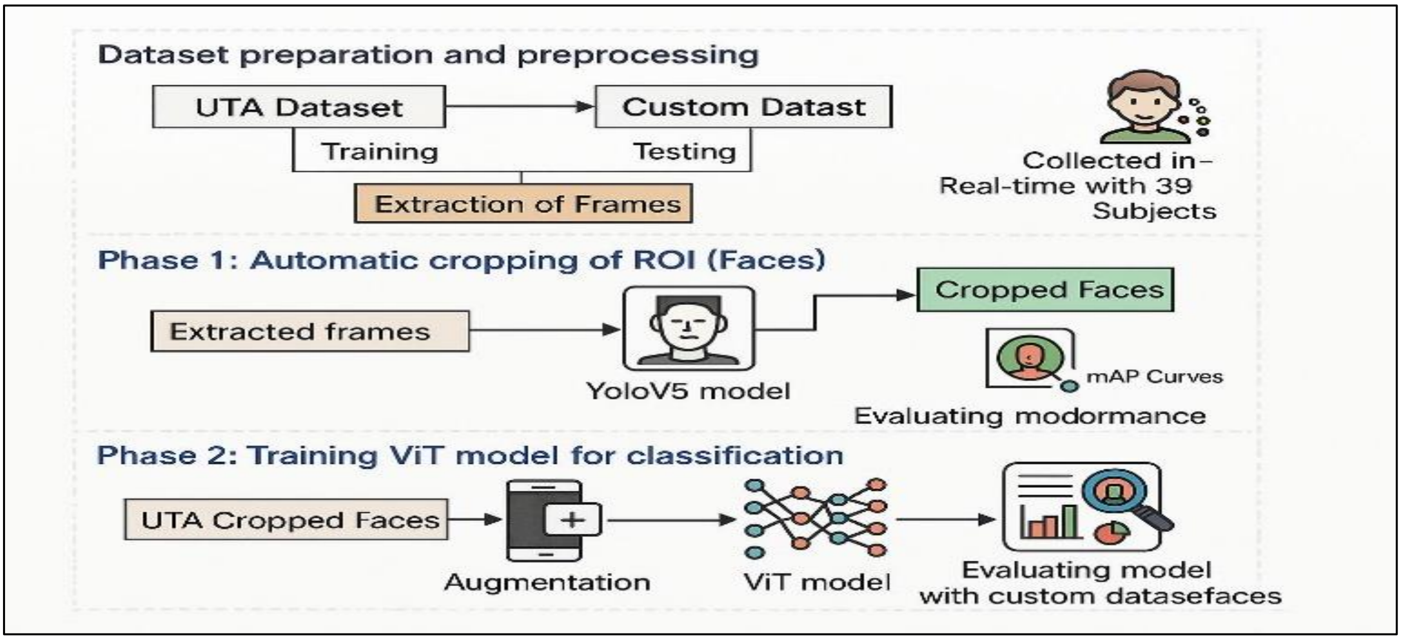


Fig 1 Proposed framework

The system utilizes two video datasets: (1) the Arlington Real-Life Drowsiness Dataset (UTA-RLDD), and (2) a custom dataset collected at Texas University. Both datasets are processed through the same framework for detecting signs of drowsiness. Video frames are extracted and fed into the pipeline as input.

➢ *Face Detection with A. Yolov5*

Since the video frames from both datasets have wide-angle views, it is essential to focus on specific regions of interest (ROIs) to improve detection accuracy. YOLOv5 is employed for this purpose, enabling automated face detection and cropping from the broader image. The YOLOv5 architecture integrates Cross Stage Partial Networks (CSP) and a variant of the Darknet backbone. These elements enhance feature extraction and reduce computation without sacrificing accuracy.In our setup, the YOLOv5 model is trained and fine-tuned to detect faces within the extracted wide-angle frames. CSP and the Darknet grid mechanism annotate the input image to extract meaningful features and recover target-related information. During the face detection phase, the input frame is divided into an A × A grid. If the center of a face lies within a specific grid cell, that cell becomes responsible for detecting the object.The confidence score for the j-th bounding box in the i-th grid cell is calculated using Equation (1):

$$Cij = Pi,j \times IOUPredictedTrue \qquad (1)$$

Here, Cij C_{ij} Cij represents the confidence score for the j j j-th bounding box in the i i i-th grid. The term Pi,j P_{i,j} Pi,j indicates whether a target is present in the j j j-th bounding box of the i i i-th grid, taking a value of 1 if a target is detected and 0 otherwise. The parameter IOUPredicted,True IOU_{Predicted,True} IOUPredicted,True refers to the Intersection Over Union (IOU), a widely used metric that measures the overlap between the predicted and actual bounding boxes. A higher IOU score indicates greater accuracy in predicting the box's location.

➢ *Image Augmentation*

Training Vision Transformer (ViT) models on large and varied image datasets plays a key role in boosting their accuracy and overall performance. To achieve this, image augmentation techniques are applied to create diverse versions of existing images. These transformations—such as flipping, shifting, zooming, and others—help the model generalize better by introducing subtle variations. This approach enhances the model's ability to recognize patterns under different conditions.

➢ *Vision Transformers for Image Classification*

Once facial detection and image augmentation are completed, an effective machine learning architecture is essential for accurate image classification. In this research, we employ the Vision Transformer (ViT) model to carry out this task efficiently. Unlike traditional convolutional neural networks, the ViT architecture leverages the power of transformer models—commonly used in natural language processing—to handle visual data. Instead of using convolutional layers, the input image is first resized and then divided into **N** fixed-size patches. Each patch is then flattened and linearly embedded to form a sequence of inputs, which are fed into the transformer model for classification.

$$I\,{}^{\epsilon}RH\times W\times C \Rightarrow I_P\,{}^{\epsilon}RN\times(P^2.C) \qquad (2)$$

$$N = H \times W/P^2 \qquad (3)$$

In Equations (2) and (3), **H**, **W**, and **C** denote the height, width, and number of channels of the original image, respectively. The term **(P, P)** refers to the resolution of each individual patch, while **N** represents the total number of patches generated from the image. An example of the patch division process applied to a sample image is illustrated in Figure 3. Further details about the characteristics of these patches are summarized in Table II.

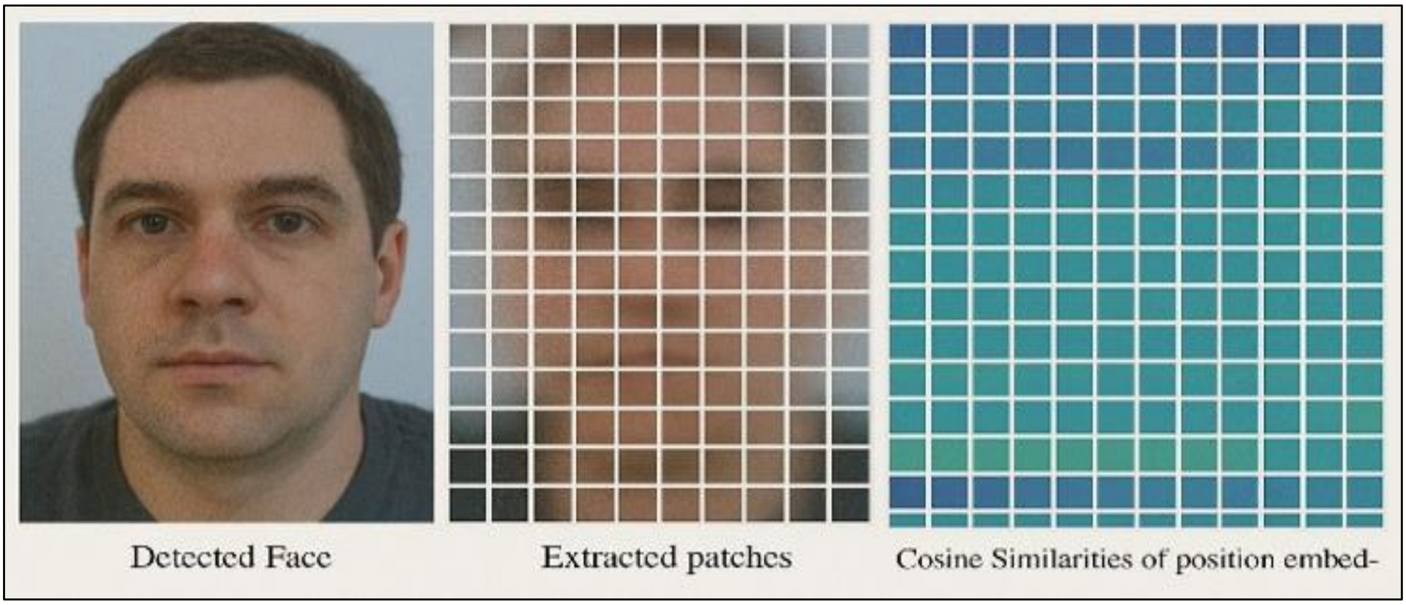


Fig 2 Patch Generation and Embedding Visualization

Table 1 Characteristics of the Created Patches

Technique	Skewnes
Size of image 392×392	
Size of patch 28×28	
Patches per image 196	
Elements per patch 2352	

Each image patch of size **(P, P)** is first flattened into a one-dimensional vector of shape **(1, P^2)**. These flattened vectors, denoted as **E(1, P^2)**, are then passed through a fully connected layer represented by the projection matrix **F(P^2, D)**. This transformation converts each patch into a fixed-length latent vector of dimension **D**, referred to as a *patch embedding*, represented as **e(1, D)**. The conceptual workflow of generating these embeddings, along with the architecture of the Vision Transformer (ViT), is depicted in Figure 4(a).

To enable classification, a learnable *class token* is added to the sequence of patch embeddings. Since images do not inherently carry positional context like text, *learnable positional embeddings* (**s**) are also introduced. These are added alongside the class token iclass i_{class} iclass to retain spatial relationships among patches. The final input sequence fed into the transformer, denoted by $\mathbf{z_0}$, is formulated in Equations (4) and (5).

*xn*E] + E*pos* (4)

Where

$$E\in \mathbb{R}^{(P^2C)\times D}, \mathbf{E}_{pos} \in \mathbb{R}^{(N+1)\times D} \quad (5)$$

These constructed sequences serve as inputs to the transformer encoder, which consists of **L** identical layers. Each layer includes a Multi-Head Self-Attention (MSA) block followed by a Multi-Layer Perceptron (MLP) block, as illustrated in the figures. The transformer operates using both of these blocks in tandem, with each encoder layer incorporating a Layer Normalization (LN) step followed by residual skip connections. The mathematical expressions for these operations are described in Equations (6) and (7).

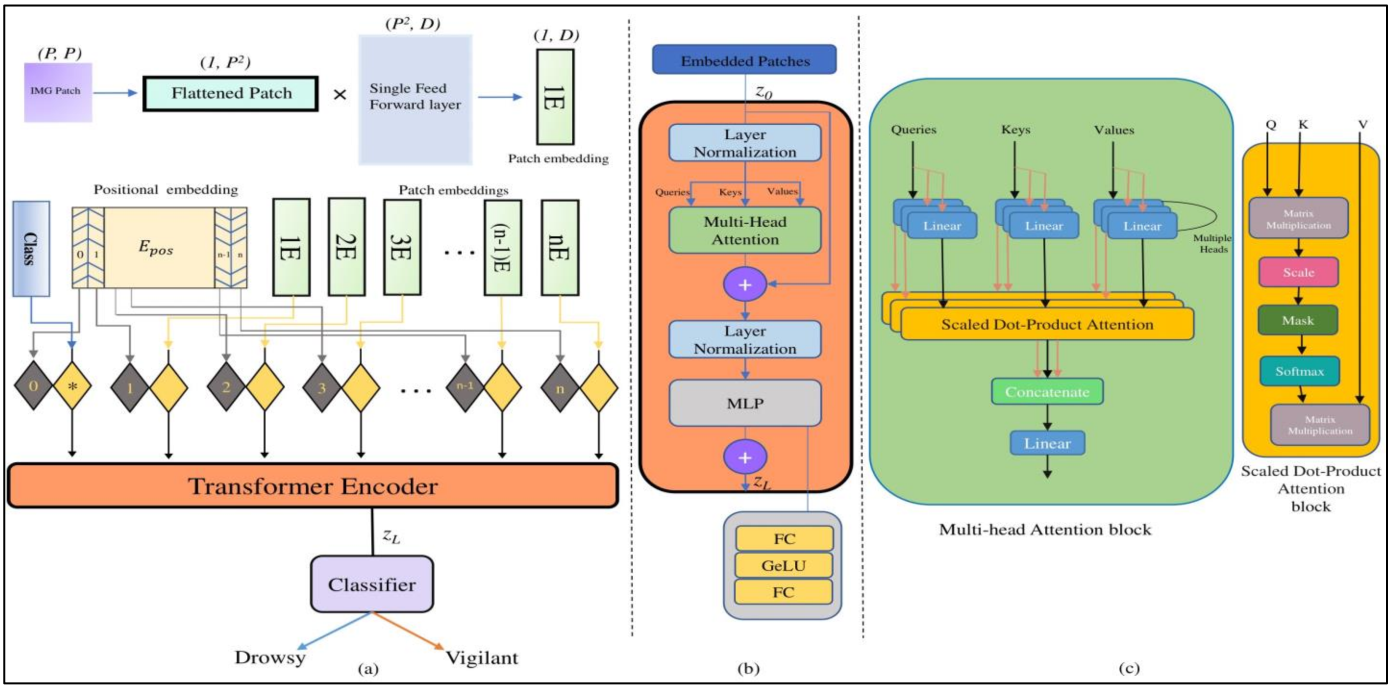


Fig 3 (a) Generation of Patch Embeddings and Conceptual Overview of the ViT Model (b) Transformer Encoder Architecture (c) Multi-Head Attention Block

$$z'_{l} = MSA(LN(z_{l-1})) + z_{l-1} \quad (6)$$

$$z_{l} = MLP(LN(z'_{l})) + z'_{l} \quad (7)$$

The operations carried out within the Multi-Head Self-Attention (MSA) block are illustrated in Figure 4(c). In the attention mechanism, the input vector is first duplicated and then multiplied with three distinct weight matrices — WqW_qWq, WkW_kWk, and WvW_vWv — resulting in the generation of the query (Q), key (K), and value (V) matrices. To compute the attention matrix, a dot product is performed between each query vector qqq in Q and all corresponding key vectors kkk in K. This is typically done by multiplying matrix Q with the transpose of matrix K. In the Self-Attention (SA) mechanism, the scaled dot-product is a variation of the regular dot-product, adjusted by dividing by the square root of the dimension of the key vectors, dkd_kdk, to stabilize gradients. The resulting scores are then passed through a softmax function to determine the attention weights. These weights are then multiplied by the value vectors to compute the final output for each attention head, as described in Equation (8). Within the transformer encoder, the MSA block performs this scaled dot-product attention independently across multiple attention heads. The outputs from all heads are concatenated and sent through a fully connected feed-forward network wi parameters.

$$SA = Softmax(\frac{QK^T}{\sqrt{d_k}}) \times V = W_{attention} \times V \quad (8)$$

$$MSA = concat(SA_1;SA_2;.....SA_h) \times W^0 \quad (9)$$

$$W0 \in Rhdk \times D$$

The MLP (Multi-Layer Perceptron) block includes a fully connected feed-forward layer that incorporates non-linear activation functions. At the final layer of the encoder, the most significant token from the sequence, denoted as zL, is sent to an external classification head, which is responsible for predicting the corresponding class labels.

IV. EXPERIMENTAL ANALYSIS AND RESULTS

➢ *Dataset Preparation and Utilization*

This study utilizes the UTA-RLDD dataset [26], a widely recognized and comprehensive benchmark dataset for drowsiness detection, to train the Vision Transformer (ViT) model. The UTA-RLDD training set includes data from 36 participants, captured under various real-world conditions. The video recordings in this dataset focus on two primary categories: drowsiness-related behaviors (such as yawning and head nodding), and non-drowsiness activities (like speaking, laughing, and looking around), with each clip lasting approximately one and a half minutes. Random frames are extracted from each participant’s video and are labeled as either ‘alert’ or ‘drowsy’ based on their visible state. These frames have a resolution of 640 × 480, which is notably higher than what most other drowsiness detection datasets offer. Additionally, the dataset features considerable variations in facial scale, orientation, and expression, making it well-suited for evaluating mode l performance in practical scenarios. This image-based dataset is used for both annotation and training of the YOLOv5 and ViT models.

For evaluation purposes, a custom dataset was created, involving 39 individuals recorded using a high-resolution DSLR camera. Compared to the UTA-RLDD dataset, this custom collection introduces greater variability in body posture, camera angles, and facial orientation. The detailed specifications of both the UTA-RLDD and the custom datasets are presented in Table III.

Table 2 Comparision of the Datasets

Attribute	**UTA-RLDD Dataset**	**Our Dataset**
Frame resolution	640×480	3840×2160
Number of Subjects	36	39
Collected in day and night	×	X
Multi-oriental frames	×	X
Number of scenarios	Five	Nine
Number of frames	9180	1246
Utilization	Training, Validation	Testing

- *Computation Specifications of Proposed System*

This section outlines both the hardware and software configuration details used in the Driver Drowsiness Detection Framework. The implementation was carried out using Python 3.9, incorporating libraries such as TensorFlow 2.0, Keras, and OpenCV for frame processing and model development. The YOLOv5 and ViT models were trained without relying on high-performance Graphics Processing Units (GPUs). A detailed overview of the system requirements and specifications necessary for training and testing the models is provided in Table IV.

Table 3 Specification and Configuration

Specifications	System's Configuration
Operating system	Ubuntu 20.04.3 LTS
CPU	Intel® i7 10th gen
RAM	15.8 Usable
GPU	Intel® UHD Graphics
Frameworks	Tensorflow, OpenCV

- *Evaluating Yolov5 for Face Detection*

YOLOv5 serves as the primary architecture in this framework, configured specifically for precise facial detection and feature extraction from input images. The model was trained using custom parameters over 200 epochs. Following extensive testing, the model achieved a minimum confidence score of 0.75. Key performance metrics such as accuracy, recall, and mean Average Precision (mAP) at a 0.5 threshold are illustrated in Figure 6. The detection results from the trained architecture, evaluated on a sample of the dataset, are presented in Figure 5. After multiple evaluations, the inference speed of the trained YOLOv5 model was determined to approximately 51.9 frames per second. Post-detection, the region of interest (ROI) is extracted by cropping the bounding box from wide-angle frames in both training and validation datasets.

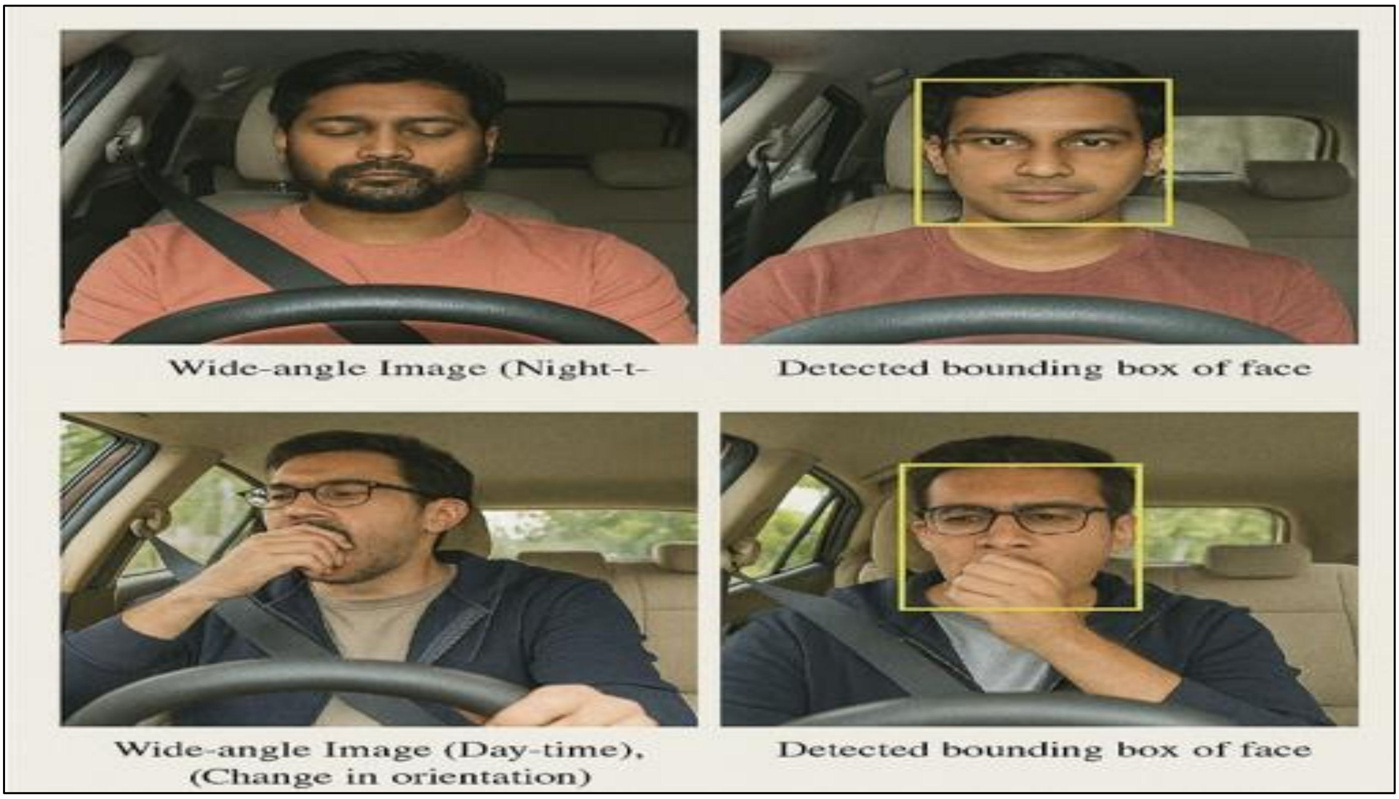

Wide-angle Image (Night-t-
Detected bounding box of face
Wide-angle Image (Day-time),
(Change in orientation)
Detected bounding box of face

Fig 4 YoloV5 results on sample images in custom dataset

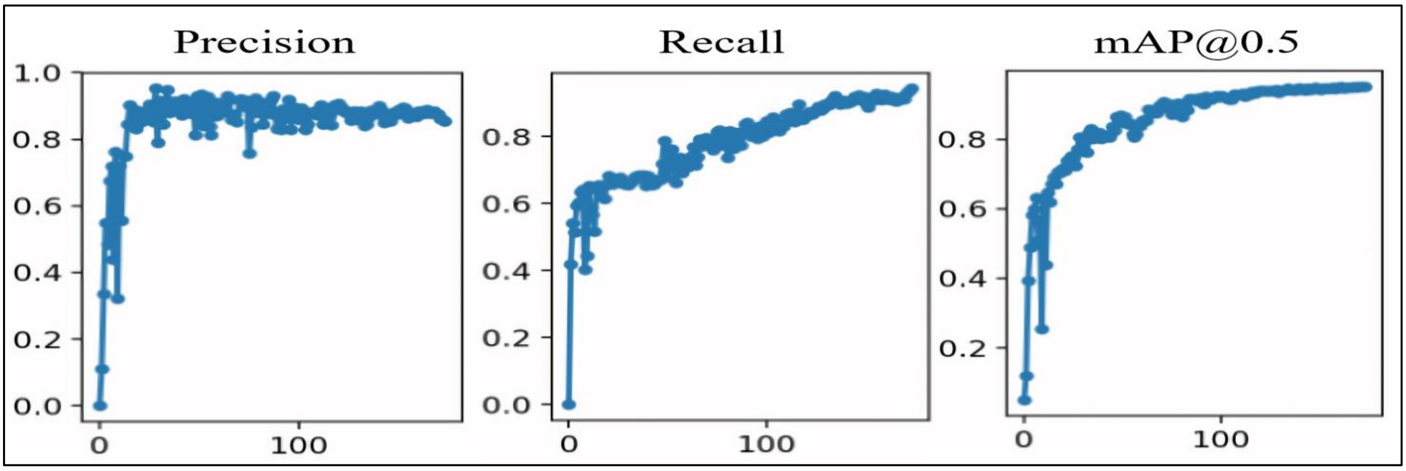


Fig 5 Visualization of YoloV5 Performance for 200 epochs

➢ *Evaluating Trained Vit Architecture*

After face detection, the training kit is improved with effective methods, as discussed in section III. In this experiment, the VIT model is trained for drowsiness, the state classification is a binary like Dowy or awake.

Specific Benchmarks for evaluating classification models are used in the evaluation of the Vic framework. The status of accuracy and loss of accuracy for model training and verification is depicted in FIG. 7. These learning plots show a good suitable algorithm because both confirmation and training decreases maintain a point of stability with minimal differences. To maximize the performance, the training of skilled VIT models was included at the same time as three assignments: 1) calculation of the output, 2) troubleshooting errors and 3) setting the hyperparameters. After several relapse of the setting of hyperparameters, the maximum training and confirmation accuracy, respectively, is respectively 96.2% and 97.4% achieved with a specific set of hyperparameters, as shown in the Table V.

Table 4 Tuned Hyperparameters of Vit Model

Hyper-parameter	Attribute
Number of classes	2
Input shape	(256, 256, 3)
Resized image Size	(392, 392)
Patch Size	28
Batch Size	256
Number of Epochs	150
Learning Rate	0.001
Weight Decay	0.0001
Number of Heads	4
Transformer Layers	8
Transformer Units	[128, 64]
MLP head Units	[2048, 1024]

To further evaluate the classification performance of the ViT model, both Hamming Loss and Binary Cross-Entropy were computed. The trained ViT model achieved a cross-entropy loss of 0.6907 and a Hamming Loss of 0.0673. Since lower values indicate better performance, the low log loss suggests promising results. While cross-entropy effectively penalizes incorrect predictions—making it a suitable choice for a loss function—it is not ideal as a standalone evaluation metric. Therefore, additional accuracy-based metrics such as Precision, Recall (also known as Sensitivity), and F1-Score were calculated to provide a more comprehensive performance overview. These metrics are derived using Equations (10), (11), and (12), respectively.

$$Precision = \frac{T_P}{T_P + F_P} \tag{10}$$

$$Sensitivity = \frac{T_P}{T_P + F_N} \tag{11}$$

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

➢ *The Influence of Training and Validation Data Splits On Test Accuracy*

The primary objective was to evaluate the performance of the ViT model's predictions on our custom dataset using various training and validation data splits. For this purpose, four different data partition ratios were applied to the UTA-RLDD dataset: 80-20, 70-30, 60-40, and 50-50. The ViT model was trained and validated on each of these splits, and the corresponding accuracy scores were compared to assess performance.

Notably, significant variations in accuracy were observed depending on the ratio of training to validation data, which in turn influenced the model's performance on the custom test dataset. Among all configurations, the 80-20 split produced the most favorable results, achieving training, validation, and test accuracies of 96.2%, 97.4%, and 95.5%, respectively, as illustrated in Figure 8.

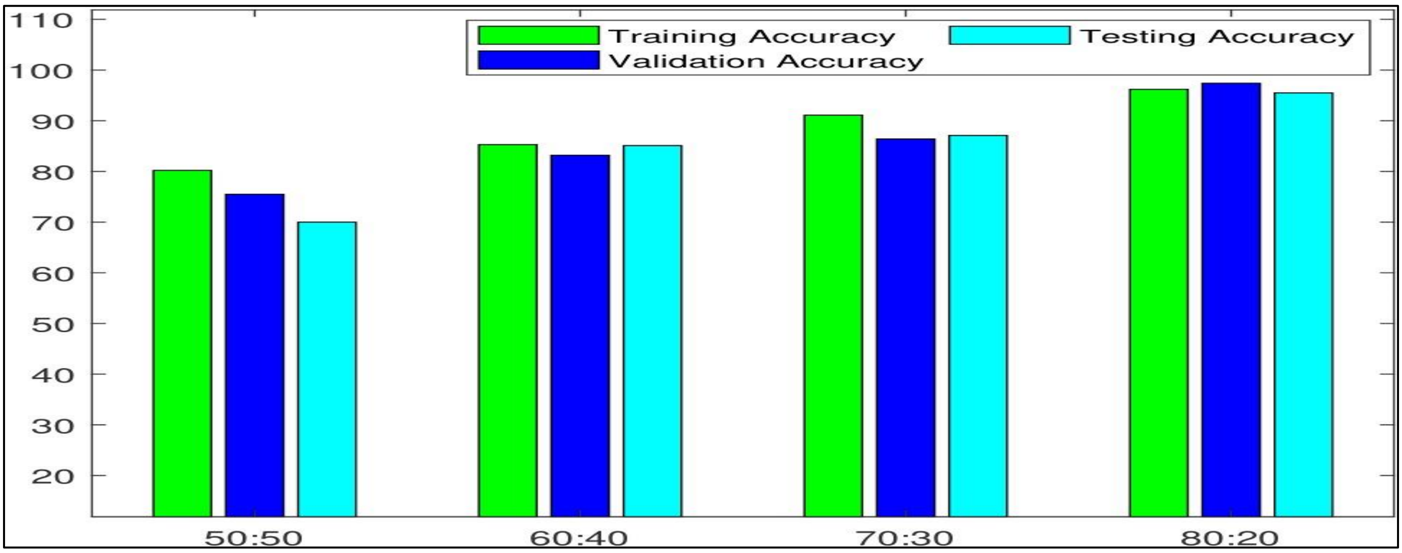
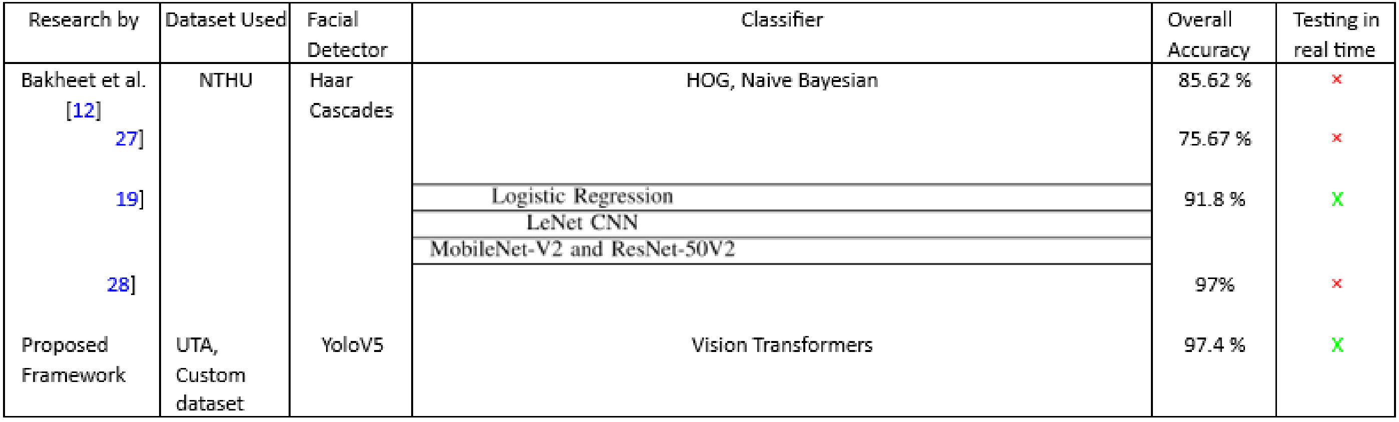


Fig 6 The Impacts of Data Splits

Table 5 Statistical Information of Performance of Vit Model with Custom Dataset

		Day-Time		Evening-Time			Night-Time		
Scenario	Drowsy (F)	Vigilant (F)	Accuracy	Drowsy (F)	Vigilant (F)	Accuracy	Drowsy (F)	Vigilant (F)	Accuracy(F)
Bare Face	0.979	0.978	0.979	0.985	0.979	0.982	0.949	0.954	0.951
Spectacles	0.955	0.959	0.957	0.989	0.979	0.984	0.892	0.917	0.905
Sunglasses	0.932	0.955	0.943	0.965	0.970	0.967	-	-	-
Average	0.955	0.964	0.959	0.980	0.976	0.978	0.921	0.935	0.928

Table 6 Comparision of Our Proposed Framework with the Existing Models



Research by	Dataset Used	Facial Detector	Classifier	Overall Accuracy	Testing in real time
Bakheet et al. [12]	NTHU	Haar Cascades	HOG, Naive Bayesian	85.62 %	×
27]				75.67 %	×
19]			Logistic Regression	91.8 %	X
			LeNet CNN		
			MobileNet-V2 and ResNet-50V2		
28]				97%	×
Proposed Framework	UTA, Custom dataset	YoloV5	Vision Transformers	97.4 %	X

➢ *Evaluation of Vit Model with Custom Dataset*

The custom dataset we created includes three different conditions: bare face, glasses, and sunglasses, with various gestures recorded during the daytime, nighttime, and early morning hours. The proposed system was tested using this custom dataset for evaluation. A statistical analysis of the model's performance across all these scenarios was conducted separately, as shown in Table VII. Notable differences were observed in the average accuracies of the model for images taken during the daytime, night, and midnight. The model achieved average accuracies of 95.9%, 97.8%, and 92.8% for daytime, evening, and nighttime images, respectively. The model performed better during the evening due to optimal lighting conditions. The final overall accuracy of the ViT model on the custom dataset was 95.5%.

➢ *Comparison with Existing Models*

An effective combination of architectures for both face detection and classification is critical for achieving optimal performance. To attain the best results, several frameworks employing different computer vision and machine learning architectures have been proposed. This section focuses on reviewing prominent studies that use machine learning architectures to detect human drowsiness. A detailed comparison of various methods for detecting drowsiness in humans is presented in Table VIII. Numerous face detection algorithms, ranging from Haar cascades to Convolutional Neural Networks (CNNs), have been developed in recent research efforts on drowsiness detection. These studies have also utilized a range of image classification algorithms, from Bayesian classifiers to CNNs.

V. AUTHOR CONTRIBUTIONS: KEY FINDINGS AND COMPARATIVE ANALYSIS

This section presents a comparative analysis of our proposed model with existing research in the field, highlighting significant findings. The comparison focuses on two key aspects: datasets and model architecture.

➢ *Dataset*

The effectiveness of drowsiness detection is highly dependent on several factors within the dataset. A large volume of data, while beneficial, can make real-time evaluation more challenging【26】【29】. This issue is compounded by the variability and ambiguity of image data in public datasets. To address these challenges, we created a custom dataset tailored to real-time requirements, which resulted in optimal test outcomes. This dataset is not only easier to analyze but also has the potential for further expansion, allowing other researchers to build on the system for global drowsiness detection.

➢ *YOLOv5 and Vision Transformer*

Many recent drowsiness detection studies have relied on Convolutional Neural Networks (CNNs)【13】, Generative Adversarial Networks (GANs)【15】, and traditional computer vision techniques【11】. In contrast, our approach introduces a novel combination of YOLOv5 and Vision Transformer (ViT) for real-time drowsiness detection, an approach that has not been implemented before. We conducted various experiments with real-time adapted datasets to assess the performance of the ViT architecture. This paper provides a comprehensive comparison between our proposed system and existing solutions, based on model performance.

VI. CONCLUSION

In this paper, we introduced the Vision Transformer (ViT) for estimating the drowsiness state of skilled drivers. The proposed system consists of two main components: the early component, which utilizes YOLOv5 for facial detection and cropping using a CNN architecture to detect five predefined facial landmarks, and the core component, which employs ViT for binary image classification. After extensive testing, YOLOv5 achieved an approximate mAP score of 95%. The ViT model demonstrated high values in key performance metrics, including average precision (0.97), sensitivity (0.98), and F1-score (0.97). Additionally, the use of a custom dataset enabled the ViT architecture to achieve an impressive test accuracy of 95.5%.

While the proposed model performs satisfactorily in terms of detection accuracy, it requires large volumes of labeled data, particularly for model training.

FUTURE WORK

There are several areas for future improvements. First, we aim to optimize the network configuration within the proposed architecture for deployment on micro-calculation systems, thereby reducing computational costs and improving efficiency without sacrificing performance. Second, we plan to use data augmentation techniques to expand the training dataset, further enhancing the model's performance.

REFERENCES

[1]. "Nhtsa report(accessed on 4 august 2021)." [Online]. Available: https://www.nhtsa.gov

[2]. J. Yu, S. Park, S. Lee, and M. Jeon, "Driver drowsiness detection using condition-adaptive representation learning framework," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 11, pp. 4206–4218, 2019.

[3]. Y. S. S. E. Viswapriya, Singamsetti Balabalaji, "A machine-learning approach for driver-drowsiness detection based on eye-state," *INTER-NATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOL- OGY (IJERT)*, vol. 10, April 2021.

[4]. M. M. Hasan, C. N. Watling, and G. S. Larue, "Physiological signalbased drowsiness detection using machine learning: Singular and hybrid signal approaches," *Journal of Safety Research*, 2021.

[5]. D. K. D. K.Mirunalini, "Drowsiness detection using deep neural network," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 9, pp. 317–326, 2021.

[6]. M. Gromer, D. Salb, T. Walzer, N. M. Madrid, and R. Seepold, "Ecg sensor for detection of driver's drowsiness," *Procedia Computer Science*, vol. 159, pp. 1938–1946, 2019, knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES2019.

[7]. M. Zhu, H. Li, J. Chen, M. Kamezaki, Z. Zhang, Z. Hua, and S. Sugano, "Eeg-based system using deep learning and attention mechanism for driver drowsiness detection," in *2021 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*, 2021, pp. 280–286.

[8]. M. Song, L. Li, J. Guo, T. Liu, S. Li, Y. Wang, Qurat ul ain, and J. Wang, "A new method for muscular visual fatigue detection using electrooculogram," *Biomedical Signal Processing and Control*, vol. 58, p. 101865, 2020.

[9]. L. R. Femilia Hardina Caryn, "Driver drowsiness detection based on drivers' physical behaviours: A systematic literature review," *Computer Engineering and Applications*, vol. 10, no. 3, 2021.

[10]. H. U. R. Siddiqui, A. A. Saleem, R. Brown, B. Bademci, E. Lee, F. Rustam, and S. Dudley, "Non-

invasive driver drowsiness detection system," *Sensors*, vol. 21, no. 14, 2021.

[11]. L. Thulasimani, P. P, and P. S P, "Real time driver drowsiness detection using opencv and facial landmarks," *Int. J. of Aquatic Science*, vol. 12, no. 2, pp. 4297–4314, 2021.

[12]. S. Bakheet and A. Al-Hamadi, "A framework for instantaneous driver drowsiness detection based on improved hog features and na¨ıve bayesian classification," *Brain Sciences*, vol. 11, p. 240, 02 2021.

[13]. A. A. Jordan, A. Pegatoquet, A. Castagnetti, J. Raybaut, and P. Le Coz, "Deep learning for eye blink detection implemented at the edge," *IEEE Embedded Systems Letters*, vol. 13, no. 3, pp. 130–133, 2021.

[14]. T. Vesselenyi, S. Moca, A. Rus, T. Mitran, and B. Tataru, "Driver˘ drowsiness detection using ANN image processing," *IOP Conference Series: Materials Science and Engineering*, vol. 252, p. 012097, oct 2017.

[15]. M. Ngxande, J.-R. Tapamo, and M. Burke, "Bias remediation in driver drowsiness detection systems using generative adversarial networks," *IEEE Access*, vol. 8, pp. 55592–55601, 2020.

[16]. Z. Li, C. Liukui, J. Peng, and Y. Wu, "Automatic detection of driver fatigue using driving operation information for transportation safety," *Sensors*, vol. 17, p. 1212, 05 2017.

[17]. S. Said, S. Alkork, T. Beyrouthy, M. Hassan, O. E. Abdellatif, and M. F. Abdraboo, "Real Time Eye Tracking and Detection- A Driving Assistance System," *Advances in Science, Technology and Engineering Systems Journal*, vol. 3, no. 6, pp. 446–454, 2018.

[18]. F. You, X. Li, Y. Gong, H. Wang, and H. Li, "A real-time driving drowsiness detection algorithm with individual differences consideration," *IEEE Access*, vol. 7, pp. 179396–179408, 2019.

[19]. R. Tamanani, R. Muresan, and A. Al-Dweik, "Estimation of driver vigilance status using real-time facial expression and deep learning," *IEEE Sensors Letters*, vol. 5, no. 5, pp. 1–4, 2021.

[20]. Z. X. Xu Q Zhu Z, Ge H Zhang Z, "Effective face detector based on yolov5 and superresolution reconstruction." *Comput Math Methods Med. 2021*, Published 2021 Nov 16.

[21]. D. Qi, W. Tan, Q. Yao, and J. Liu, "Yolo5face: Why reinventing a face detector," 2022.

[22]. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai,

[23]. T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020.

[24]. H. Dong, L. Zhang, and B. Zou, "Exploring vision transformers for polarimetric sar image classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2021.

[25]. P. Deng, K. Xu, and H. Huang, "When cnns meet vision transformer: A joint framework for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[26]. Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sensing*, vol. 13, no. 3, 2021.

[27]. R. Ghoddoosian, M. Galib, and V. Athitsos, "A realistic dataset and baseline temporal model for early drowsiness detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[28]. S. Mittal, S. Gupta, Sagar, A. Shamma, I. Sahni, and N. Thakur, "Driver drowsiness detection using machine learning and image processing," in *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2021, pp. 1–8.

[29]. A.-C. Phan, N.-H.-Q. Nguyen, T.-N. Trieu, and T.-C. Phan, "An efficient approach for detecting driver drowsiness based on deep learning," *Applied Sciences*, vol. 11, no. 18, 2021.

[30]. S.-H. L. Ching-Hua Weng, Ying-Hsiu Lai, "Driver drowsiness detection via a hierarchical temporal deep belief network," *In Asian Conference on Computer Vision Workshop on Driver Drowsiness Detection from Video, Taipei, Taiwan*, Nov. 2016.