ISSN: 1692-7257 - Volume 1 – Number 47 - 2026

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

Digital Object Identifier: 10.24054/rcta.v1i47.3824

Safe driving detection by Vision Transformer and convolutional networks comparison

Comparación de redes Vision Transformer y convolucionales para detección de conducción segura

PhD. Robinson Jiménez Moreno [1], MSc. Anny Astrid Espitia Cubillos [2], MSc. Javier Eduardo Martínez Baquero [3]

[1] ***Universidad Militar Nueva Granada,*** *Facultad de ingeniería, Programa de Ingeniería Mecatrónica, Bogotá, Colombia.*
[2] ***Universidad Militar Nueva Granada,*** *Facultad de ingeniería, Programa de Ingeniería Industrial, Bogotá, Colombia.*
[3] ***Universidad de los Llanos,*** *Facultad de Ciencias Básicas e Ingeniería, Villavicencio, Colombia.*

Correspondence: *anny.espitia@unimilitar.edu.co*

Received: *July 23, 2025.* ***Accepted:*** *December 20, 2025.* ***Published:*** *January 01, 2026.*

How to cite: *R. Jiménez Moreno, A. A. Espitia Cubillos y J. E. Martínez Baquero, "Comparación de redes Vision Transformer y convolucionales para detección de conducción segura", RCTA, vol. 1, n.º. 47, pp. 62-71, Jan. 2026. Recuperado de https://ojs.unipamplona.edu.co/index.php/rcta/article/view/3824*

Abstract: This paper presents the results of comparing the training of deep learning architectures applied to the development of safe driving systems. Databases were generated with 670 images of drivers inside vehicles, which were divided into three subsets for training two architectures based on convolutional neural networks (CNNs) and transformer networks for vision. 70% of the images were used for training, 20% for validation, and the remaining 10% for testing. These two architectures were compared to assess their pattern recognition capabilities in classifying three driving states, normal state, distracted state and sleep state. In both cases, the need to focus the learning to improve the learning performance of the two architectures is evident, for which a previous stage of face segmentation by means of Haar classifier is included, obtaining accuracy levels of 98% for the CNN and 87% for the Transformers network with average inference times of 0.1 and 0.52 seconds, F1 scores of 98.9% and 82.2%, and recall rates of 98.8% and 80.6%, respectively, the statistical metrics for each class demonstrate a high degree of confidence in the recognition of each class. The comparison was performed on a computer with a 2.3GHz Core i9 processor, 24GB of RAM, and an RTX 4080 GPU with 12GB of memory, using MATLAB® programming software.

Keywords: driving assistant, convolutional neural networks, drowsiness detection, haar classifier, safe driving, transfer learning, computer vision.

Resumen: Este documento presenta los resultados de comparar el entrenamiento de arquitecturas de aprendizaje profundo aplicadas al desarrollo de sistemas de conducción segura. Se generan bases de datos con capturas de 670 imágenes de conductores en el interior del vehículo, que se dividieron en tres subconjuntos para el entrenamiento de dos arquitecturas basadas en redes neuronales convolucionales (CNN) y redes transformers para visión, el 70% de las imágenes se utilizó para el entrenamiento, el 20% se destinó a la validación y el 10% restante se reservó para las pruebas. Estas dos arquitecturas se

comparan con el fin de contrastar su capacidad en el reconocimiento de patrones en la clasificación de tres estados de conducción, estado normal, estado de distracción y estado de sueño. En ambos casos se evidencia la necesidad de focalizar el aprendizaje a fin de mejorar el desempeño en el aprendizaje de las dos arquitecturas, para lo que se incluye una etapa previa de segmentación de caras mediante clasificador Haar, obteniéndose niveles de precisión del 98% para la CNN y del 87% para la red Transformers, tiempos promedio de inferencia de 0.1 y 0.52, F1-score de 98.9% y 82.2%, y recall de 98.8% y 80.6%, respectivamente, las métricas estadísticas por clase evidencian el alto grado de confianza en el reconocimiento de cada clase. La comparativa se realiza en un equipo de cómputo con procesador core i9 de 2.3GHz y 24GB de RAM, una GPU RTX 4080 de 12 GB de memoria, bajo software de programación MATLAB®.

Palabras clave: asistente de conducción, redes neuronales convolucionales, detección de somnolencia, clasificador Haar, conducción segura, transferencia de aprendizaje, visión por computador.

1. INTRODUCTION

In recent years, ensuring safe driving [1] has become a critical concern due to the increasing number of traffic accidents worldwide [2]. Advanced driver assistance systems (ADAS) and driver monitoring systems (DMS) [3] have emerged as promising solutions to enhance road safety by detecting signs of drowsiness, distraction, or unsafe driving behavior. With the rapid advancement of deep learning, computer vision techniques have played a central role in improving the performance of these systems. Among the most prominent approaches, within the state of the art and associated with non-invasive systems such as artificial intelligence algorithms, are Convolutional Neural Networks (CNNs), which have demonstrated remarkable success in image-based tasks [4] [5], and Vision Transformers (ViTs), a more recent architecture that has shown outstanding results in various vision applications [6].

Road safety has become one of the most pressing global concerns in recent years [7]. According to the World Health Organization, road traffic accidents claim approximately 1.35 million lives each year, and an even larger number of individuals suffer serious injuries [8]. One of the major contributors to these accidents is human error, often caused by drowsiness, distraction, or risky driving behaviors. To address this challenge, advanced technologies have been developed to monitor driver behavior and detect signs of unsafe driving, aiming to prevent accidents before they happen. Among these technologies, computer vision-based systems [9] have gained significant attention due to their ability to process visual data in real-time and provide accurate assessments of driver status.

In the field of computer vision, deep learning has emerged as a transformative approach that enables models to learn complex patterns and features directly from image-based data. Two leading architectures in this field are CNNs and ViT [10] [11]. CNNs have been the cornerstone of image classification and object detection tasks for over a decade, demonstrating impressive performance across a wide range of applications. Their hierarchical feature extraction, which leverages local connectivity and weight sharing, makes them particularly well-suited for processing images and videos. However, CNNs have certain limitations, such as their reliance on large, labeled datasets and difficulties in capturing long-range dependencies within an image.

In contrast, Vision Transformers represent a newer approach that applies the Transformers architecture, originally designed for natural language processing tasks, to vision problems [12]. ViTs divide images into patches and process them as sequences, enabling the model to capture global context through self-attention mechanisms. This design allows Vision Transformers to overcome some of the shortcomings of CNNs, particularly in modeling long-range relationships and learning more holistic representations. Recent studies have shown that ViTs can outperform CNNs on various image recognition benchmarks, provided they are trained on sufficiently large datasets.

Researches oriented to safe driving systems are largely focused on autonomous technologies, particularly object detection tasks such as pedestrian recognition and obstacle avoidance [13] [14]. However, research centered on driver-focused safety systems remains highly relevant and continues to advance. These efforts address critical

areas such as monitoring driver sobriety [15], assessing driving behavior through imbalance or erratic movement detection [16], and, most notably, detecting signs of driver drowsiness or sleepiness [17], where, for example, sleep detection requires a clear identification of eye opening [18].

While autonomous systems are designed to minimize human error, driver-centered approaches remain crucial for enhancing road safety. By continuously monitoring the driver's physical and cognitive state, these systems can provide timely alerts and help prevent accidents caused by fatigue, distraction, or other human factors. Instead of replacing the driver, they work alongside, offering support when needed and reducing risks on the road. When combined, autonomous technologies and driver-focused strategies create a more complete and effective safety framework. This integrated approach not only addresses technical challenges but also accounts for the human element, ultimately contributing to safer and more reliable driving environments for everyone on the road.

New learning models are being developed to improve driver sleep detection, using approaches such as hybrid networks [19] and electroencephalographic (EEG) signal capture [20] [21]. EEG has also been applied to detect driver fatigue [22], often combined with advanced feature extraction methods like wavelet analysis [23] and fuzzy logic [24]. More recently, deep learning algorithms have demonstrated strong performance in this area [25], successfully working with both EEG signals [26] and in identifying anomalies in driving behavior or trajectory [27]. These developments highlight the growing role of machine learning and signal processing in enhancing driver monitoring systems and improving overall road safety, with the limitation of requiring the capture of the patient's EEG signals.

Among the main deep learning algorithms are the CNN convolutional neural networks [28], which have also proven to be efficient in the detection of sleep-in drivers based on ResNet architectures [29] [30]. In this case, there are pre-trained models by transfer learning [31], using robust CNN architectures such as the YOLO network [32], based on visual identification.

The developments presented have demonstrated the advantages of deep networks in drowsiness detection; however, these works involve human intervention (EEG capture) or specificity of eye detection, which limits their applications in real time and scenarios.
More recently in the state of the art, deep learning models such as short- and long-term memory networks are used [33] and for image detection, Transformers networks are gaining strength [34], which are also beginning to be validated in autonomous driving such as traffic signal detection [35]. However, its advantage over CNN algorithms aimed at sleep detection in a safe driving environment is not clear at the time of the literature review.

In line with the research presented and the advantages of deep learning algorithms for safe driving, this work presents a comparison between a convolutional network architecture and a pre-trained ViT model [36] [37] for detecting driving states classified as normal, distracted, or sleepy. By evaluating the performance of these two approaches, the study contributes to the state of the art in driver monitoring systems, providing response times in the inference of each network in real driving scenarios, under non-invasive and eye-centric systems, which gives more generality to the learning of the fatigue pattern.

The comparison aims to highlight the strengths and limitations of each model, particularly regarding their robustness and ability to generalize across various driving conditions. This analysis provides valuable insights into which architecture may be better suited for real-world applications, helping to inform the development of more reliable and efficient systems for enhancing driver safety and reducing road accidents.

This article is divided into four sections, the introduction with an exposition of the state of the art and the objective of this work. The methodology, where the characteristics of the database and the architectures used are exposed. The analysis of results, where the performance and classification characteristics are shown, and finally the conclusions are reached.

2. METHODOLOGY

The proposed methodology, based on applied research, aims to establish a database under real-world driving conditions with different drivers. Since the state-of-the-art reports results from CNN-based architectures such as ResNet or YOLO, a proprietary CNN architecture is proposed for comparison. Using the same database, transfer learning is employed with the ViT architecture to

obtain metrics such as accuracy levels, inference time, F1 score, and recall. Based on the number of learning parameters, the network size, which impacts memory usage in a real-world application, is also analyzed. Table 1 illustrates the software and hardware characteristics used. Finally, the results are presented in accordance with some of those reported for ResNet and YOLO architectures.

Table 1: *Software and hardware configuration*

Software	Programming environment	MATLAB
	OS	Windows 11
Hardware	CPU	Intel core i9 2.3GHz
	GPU	RTX 4080
	RAM	CPU 24GB/GPU 12GB

To evaluate the performance of convolutional and transformer networks in identifying states relevant to safe driving, a user database was built across three distinct scenarios. The first scenario, labeled as "normal," represents driver's attention directed straight ahead toward steering wheel and road. The second scenario, "distraction," captures moments when driver's eyes are diverted, causing a loss of focus on road environment. Finally, "sleep" state is characterized by driver either having closed eyes or a downward-tilted head, indicating drowsiness or microsleep episodes. This database enables assessment of how effectively each network can distinguish between these critical conditions to enhance driving safety systems.

Fig. 1 shows part of the database used, showing the states of sleep, distraction and normal driving with different users behind the wheel. The complete database consists of 670 images of ten test subjects, which were divided into three subsets for model development: 70% of the images were used for training, 20% were allocated for validation, and the remaining 10% were set aside for testing. This distribution ensures the models are effectively trained, fine-tuned, and evaluated on separate data. The database is built within a daytime lighting range from 6 am to 6 pm, where each capture is made for a balanced distribution of each of the three established classes, ensuring that the same pose per user is not repeated in the distribution of the subsets. The diverse user representation in the database helps improve the generalization and robustness of the models when applied to real-world driving scenarios.

Fig. 1. *Initial database extract*

In the case of the convolutional network (CNN), the architecture illustrated in Table 2 is used. There, the following references are used for the feature extraction stage, where the structure of each learning kernel (N) has the following notations: C convolution, B Normalization Batch, R for the linear rectification unit Relu, P for the pooling dimensionality reduction layer and for classification stage FC as Fully connected. The learning kernel by convolutional kernel (N) is composed of square filters of side L and number of filters D under the L/D ratio in Table 2. In turn, the Maxpooling filter size (M), the padding (P) and the step or stride (S) are defined, the step of the maxpooling operation is kept at 1. The classification stage uses Dropout at 50% and linear activation functions (RELU).

Table 2: *CNN Architecture*

Layer	Structure	Kernel	M/P/S
N1	C-B-R	15/12	0/2/1
N2	C-B-R-P	5/24	[3 2]/2/1
N3	C-R-P	5/48	2/1/1
N4	C-R-P	3/48	2/1/1
N5	C-R-P	4/96	2/0/1
N6	C-R-P	4/96	2/0/1
N7	C-B-R	[3 4/192]	0/1/2
FC	1024-2048-3		

For the case of pre-trained ViT network, this is based on transfer learning under model presented in [36] [37] [38]. This model has 143 layers, where the input image is handled using 16 patches. Fine-tuning is performed during network training by freezing all layers except the attention layer and modifying the output layer to match the target classes. The GELU activation function and a 10% dropout are used. A data augmentation technique based on rotation and reflection of the initial database is employed.

Each network is trained with the same final parameters which are shown in Table 3. It is important to note that the input volume is conditioned by the ViT network through transfer

learning to a square image with sides of 384 pixels. Preprocessing is performed during the resizing of the database images to maintain the aspect ratio of the original image.

Table 3: *Training parameters*

Parameter	CNN	ViT
Input	384x384x3	384x384x3
Learning Rate	0.00001	0.0001
Epochis	80	50
Minilot	12	4
Optimizer	ADAM	ADAM

4. RESULTS

Following the training process, the accuracy graph initially indicates superior performance by the CNN network, as shown in Fig. 2. CNN achieves an accuracy of 91%, significantly outperforming the Vision Transformer (ViT) network, which reaches only 64.4% accuracy, as shown in Fig. 3. This notable difference highlights CNN's possible advantage in learning key features at the early stages of training. The results suggest that CNN's architecture is better suited for capturing the relevant patterns required for this task, while the ViT network may require further fine-tuning or larger datasets to improve its performance.

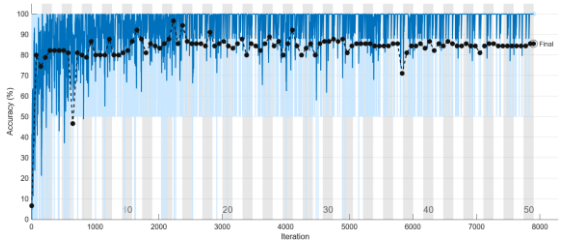

Fig. 2. *CNN network initial training graph*

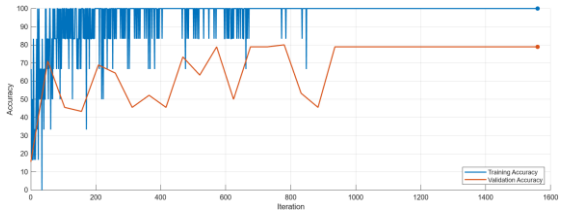

Fig. 3. *ViT network initial training graph*

However, when analyzing the confusion matrix of the CNN network (Fig. 4), it is evident that the network does not discriminate well between the three classes, eliminating the sleep class, which is attributed to the fact that given the change of scale in the image, the identification between closed and open eye is not possible. Table 4 illustrates the performance obtained by class, for which ViT exhibits better behavior by identifying something from each class.

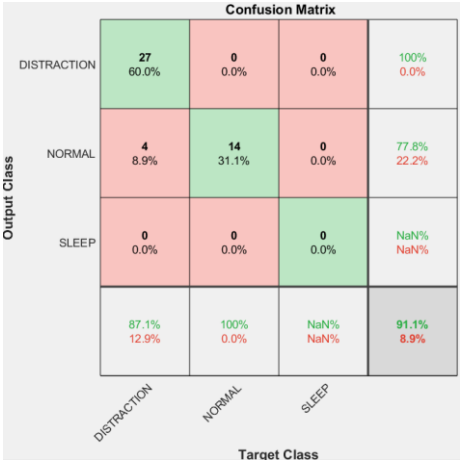

Fig. 4. *CNN network initial confusion matrix*

Table 4: *Validation summary by class*

Class	CNN(%)	ViT(%)
Distraction	60	31
Normal	31.1	23.4
Sleep	0	10

Derived from these results it is determined to employ a Haar classifier for face recognition [26], applied to each initial image, thereby generating a new set of images for training. Fig. 5 provides an excerpt from the updated database, where the top row depicts images of the sleep state, the middle row shows images of the distracted state, and the bottom row represents images of the normal driving state. This approach enhances the dataset by focusing on key facial features, allowing for more accurate classification of driver states. The refined database is then used to train models for detecting driver conditions, improving the performance and robustness of the system.

Fig. 5. *Face database extract*

After retraining the networks using the same parameters listed in Table 3 and the face database, we achieved performance results with an accuracy

of 98% for the CNN network and 87% for the ViT network.

Fig. 6 and Fig. 7 display the performance results for each respective network. These results highlight CNN's superior accuracy, showcasing its effectiveness in detecting driver states, while the ViT network, though slightly less accurate, still demonstrates strong performance. Both networks show promise in driver monitoring applications, with the customized CNN proving to be more efficient in this case.

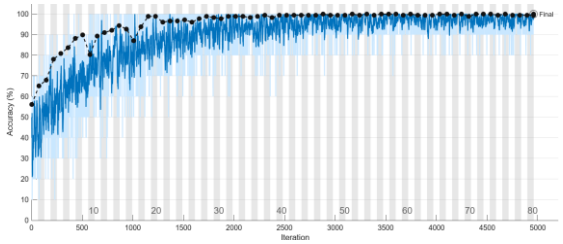

Fig. 6. *CNN network face training graph*

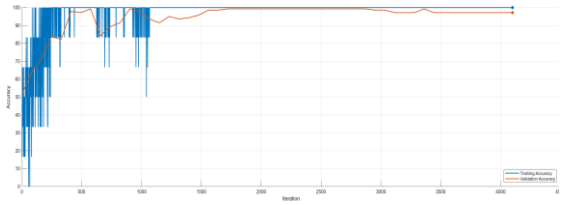

Fig. 7. *ViT network face training graph*

Fig. 8 shows the confusion matrix obtained for CNN network and Fig. 9 confusion matrix obtained for ViT network. These results are tabulated in Table 5 and show the improvement in the outcome for each class for CNN. A strong confounding effect is observed in the CNN in the distraction class, mainly with sleep. This behavior is also evident in the ViT but in a higher proportion, this class is the one that generates the main difference between the recognition of each type of network. In the case of safe driving and sleep detection, both networks manage to identify a high percentage of this category, where the CNN achieves 100% recognition, being fundamental.

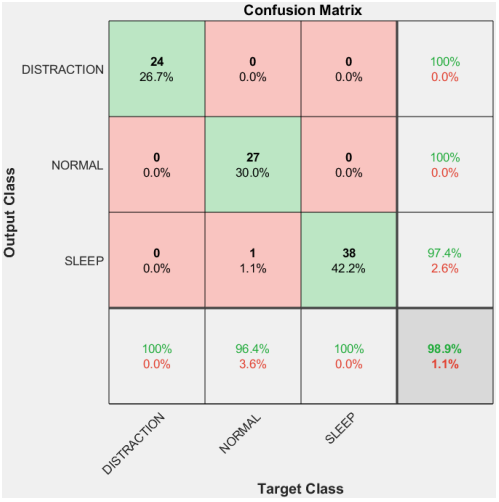

Fig. 8. *CNN network confusion matrix*

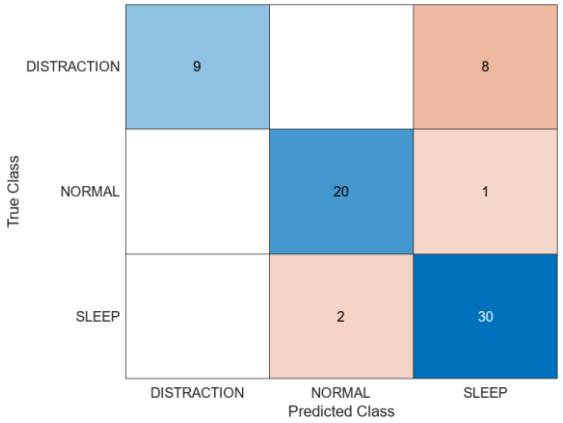

Fig. 9. *ViT network confusion matrix*

Table 5: *Final validation summary by class*

Class	CNN(%)	ViT(%)
Distracction	26.7	22.8
Normal	30	26.2
Sleep	42.2	38.1
Precision	98.9	87.1
F1-score	98.9	82.2
Recall	98.8	80.6

Table 6 allows the results of both networks to be shown graphically. CNN's bias toward the distraction class and its confusion between sleep and distraction. In contrast, the ViT model correctly identified each class without such confusion. Notably, an image showing a pronounced yawn—absent from the training dataset—was tested, and both networks consistently classified it under the distraction category. This example highlights the CNN's tendency to misclassify similar behaviors, while the ViT demonstrated stronger generalization. The consistent agreement on the yawn image also underscores the challenges posed by unseen data

and ambiguous facial expressions in classification tasks.

Table 6: *Graphical validation results*

Net	Results
CNN	
ViT	

Similarly, extreme operating conditions of the algorithm were analyzed, which can be seen in Fig 10. Strong lateral tilts or turns of the face are delimited by the operation of the Haar classifier, for which the detection of the network was conditioned to the last recognized value in order to avoid a false detection state, after 10 frames an alarm is generated as a fourth state called "no detection".

Fig. 10 shows some detection errors in the sleep category, due to factors such as eye occlusion (lower left), a key element in identifying this state. Similarly, as shown in Fig. 9 regarding the confusion matrix, detection errors occur in the distraction category when comparing sleep to sleep. In Fig. 10 (lower right), when looking down and only partially seeing the eye without evidence of the sclera, the system detects it as sleep.

Fig. 10. *Extreme operating conditions*

The findings show that the CNN, with just 30 layers, is nearly five times faster than the ViT, which contains 143 layers. This significant difference highlights the advantage of the CNN in terms of speed and computational efficiency, making it more suitable for real-time applications. In contrast, while ViT offers higher accuracy and better generalization, it comes at the cost of slower processing times due to its deeper architecture and more complex computational requirements with regard to training, such as a GPU with at least 12GB of RAM.

A test was conducted to evaluate the inference time of each network and to gain insight into its impact on a real-time system and the number of frames per second at which the detection system should operate. A dataset of 25 images per class was used to estimate the average inference time, and the results are summarized in Table 7. This table shows that detection could be achieved at a maximum of 10 fps using the CNN network, while the ViT network would operate too slowly for a useful system.

Table 7: *Average inference times*

Net	Average inference time	FPS Max
CNN	0.102725	10
ViT	0.528490	2

For a quantitative analysis, 2000 CNN detection confidence levels were used and tabulated in Table 8. It can be inferred that the confidence distribution by class is very high for the normal and sleep states, and to a lesser degree for the distracted state, which exhibits the highest variability at 31%. This indicates that the system is very reliable in positive detections for each class, achieving a 75th percentile with confidence levels close to 1 for sleep detection, the most critical state analyzed by the network.

Table 8: *Final Validation Statistics by Class*

Class	Distraction	Normal	Sleep
Mean	0.722	0.927	0.923
Median	0.988	1	1
Standard Desviation	0.315	0.224	0.2323
Variance	0.099	0.0503	0.0539

To validate the functionality as a fatigue detection application based on the three detected classes, 3 videos of three users with different scenarios and durations were used. An alert graph is included in the driving state detection algorithm to monitor the driver's condition over time. Fig. 11 presents the driver's state across the observation window, displaying three distinct levels determined by the network's classification output. Specifically, a threshold of 10 corresponds to a normal driving state, 50 indicates a distracted state, and 90

represents a sleep state. When the system fails to detect the driver's face, the state value drops to zero, signaling an absence of detection. This graphical representation helps visualize transitions between states and highlights critical moments when the driver's face is not recognized, providing essential insights for safety interventions and ensuring timely alerts are issued when risk levels increase.

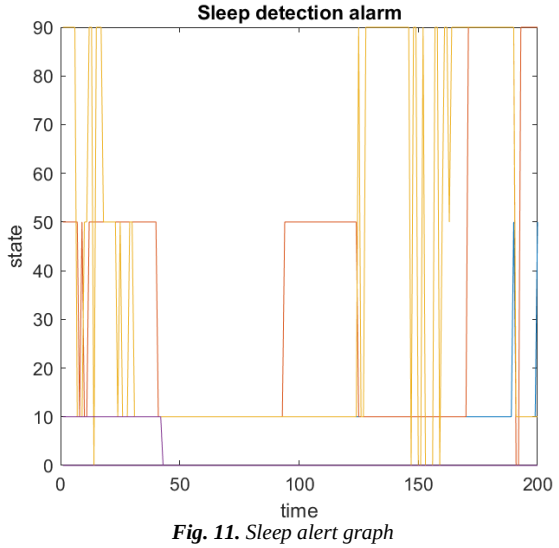

Fig. 11. *Sleep alert graph*

Under the applied algorithm, dominance metrics can be obtained for the driving task, where the driver's fatigue state is determined using equations (1), (2), and (3).

$$I_s = \frac{Fsleep}{Ftotal} \quad (1)$$

$$I_d = \frac{Fdistraction}{Ftotal} \quad (2)$$

$$I_F = \frac{Fsleep+Fdistraction}{Ftotal} \quad (3)$$

Table 9 illustrates the dominance of metrics in the real-world test videos. These results suggest that driver three (video 3) drives in a fatigued state for a longer period, so if recovery breaks are implemented for this driver, the driving times should be shorter.

Table 9: *Fatigue Dominance Metrics*

Video	Sleep	Distraction	Fatigue
1	14.5 %	10%	27.4%
2	22.31%	9.18%	31.50%
3	16.8%	12.5%	44.3%

5. CONCLUSIONS

Based on the accuracy, F1 score, and inference times, it can be concluded that a CNN network performs better for the application of detecting established safe driving states. It exhibits 11.8% greater classification accuracy than ViT and average inference times 5.14 times shorter compared to ViT.

The statistical metrics for each class of the CNN architecture demonstrate that the detection of the most critical state, corresponding to sleep, is reliable, with the inference level exceeding 0.95 for 90% of detections. While the detection of distraction states can be improved, their confusion with the sleep class continues to generate driving alerts that favor the response of the proposed system.

Future work includes conducting evaluation using public databases, as well as the possible integration of a hybrid CNN-LSTM architecture to validate results based on temporal information and explore lighter variants of Vision-Transformer architectures.

ACKNOWLEDGMENTS

The authors acknowledge the Universidad Militar Nueva Granada and the Universidad de los Llanos, where they serve as associate professors. This work is a derivative product of the research project entitled "Design of a Human–Robot Interaction Model Using Deep Learning Algorithms" (INV-ING-3971), funded by the Vice-Rectory for Research of the Universidad Militar Nueva Granada, 2024 funding period.

REFERENCES

[1] Y. Y. Wang and H. Y. Wei, "Safe Driving Capacity of Autonomous Vehicles," in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, 2018, pp. 1–5. doi:10.1109/VTCFall.2018.8690822.

[2] J. W. Lee, B. J. Park, K. H. Kim, and H.K. Choi, "A testbed for development and test of the safe driving system," in *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, 2016, pp. 1149–1151. doi:10.1109/ICTC.2016.7763392.

[3] G. Salzillo, C. Natale, G. B. Fioccola, and E. Landolfi, "Evaluation of Driver Drowsiness based on Real-Time Face Analysis," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020, pp. 328–335. doi:10.1109/SMC42975.2020.9283133.

[4] E. Karakullukcu, "Leveraging convolutional neural networks for image-based classification of feature matrix data," *Expert Syst. Appl.*, 2025, vol. 281, p. 127625, doi:10.1016/j.eswa.2025.127625.

[5] A. Abdullah, W. S. Wong, and D. Albashish, "EB-CNN: Ensemble of branch convolutional neural network for image classification," *Pattern*

Recognit. Lett., 2025, vol. 189, pp. 1–7, doi:10.1016/j.patrec.2024.12.017.

[6] Y. L. Chen, C. L. Lin, Y. C. Lin, and T. C. Chen, "Transformer-CNN for small image object detection," *Signal Process. Image Commun*., 2024, vol. 129, p. 117194, doi:10.1016/j.image.2024.117194.

[7] Y. Y. Wang and H. Y. Wei, "Road Capacity and Throughput for Safe Driving Autonomous Vehicles," *IEEE Access*, 2020, vol. 8, pp. 95779–95792, doi:10.1109/ACCESS.2020.2995312.

[8] K. Aati, M. Houda, S. Alotaibi, A. M. Khan, N. Alselami, and O. Benjeddou, "Analysis of Road Traffic Accidents in Dense Cities: Geotech Transport and ArcGIS," *Transp. Eng.*, 2024, vol. 16, p. 100256, doi:10.1016/j.treng.2024.100256.

[9] H. T. N. Le and H. Q. T. Ngo, "Application of the vision-based deep learning technique for waste classification using the robotic manipulation system," *Int. J. Cogn. Comput. Eng*., 2025, vol. 6, pp. 391–400, doi:10.1016/j.ijcce.2025.02.005.

[10] I. Shad, Z. Zhang, M. Asim, M. Al-Habib, S. A. Chelloug, and A. A. El-Latif, "Deep learning-based image processing framework for efficient surface litter detection in Computer Vision applications," *J. Radiat. Res. Appl. Sci.*, 2025, vol. 18, no. 2, p. 101534, doi:10.1016/j.jrras.2025.101534.

[11] M. Ciranni, V. Murino, F. Odone, and V. P. Pastore, "Computer vision and deep learning meet plankton: Milestones and future directions," *Image Vis. Comput.*, 2024, vol. 143, p. 104934, doi:10.1016/j.imavis.2024.104934.

[12] A. Khan, Z. Rauf, A. Sohail, A. R. Khan, H. Asif, A. Asif, and U. Farooq, "A survey of the vision transformers and their CNN-transformer based variants," *Artif. Intell. Rev.*, 2023, vol. 56, no. 3, pp. 2917–2970, doi:10.1007/s10462-023-10595-0.

[13] X. Sun, L. Jin, H. Wang, Z. Huo, Y. He, and G Wang, "Spatial awareness enhancement based single-stage anchor-free 3D object detection for autonomous driving," *Displays*, 2024, Vol. 85, p. 102821, doi:10.1016/j.displa.2024.102821.

[14] Y. Zhou, and X. Zeng, "Towards comprehensive understanding of pedestrians for autonomous driving: Efficient multi-task-learning-based pedestrian detection, tracking and attribute recognition," *Robotics and Autonomous Systems*, 2024, Vol. 171, p. 104580, doi:10.1016/j.robot.2023.104580.

[15] C. M. Farmer, "Potential lives saved by in-vehicle alcohol detection systems", *Traffic Injury Prevention*, 2021, Vol. 22, no. 1, pp. 7-12, doi:10.1080/15389588.2020.1836366.

[16] Z. Wang, Z. Li, Z. Li, Y. Xu, F. Qi, J. Kong, "A low cost and effective multi-instance abnormal driving behavior detection system under edge computing", *Computers & Security*, 2023, Vol. 132, p. 103362, doi:10.1016/j.cose.2023.103362.

[17] Y. X. Chew, S. F. Abdul Razak, S. Yogarayan, and S. N. M. S. Ismail, "Dual-Modal Drowsiness Detection to Enhance Driver Safety," *Computers, Materials and Continua*, 2024, Vol. 81, no. 3, pp. 4397-4417, doi:10.32604/cmc.2024.056367.

[18] Y. Sun, R. Wang, H. Zhang, N. Ding, S. Ferreira, and X. Shi, "Driving fingerprinting enhances drowsy driving detection: Tailoring to individual driver characteristics," *Accident Analysis & Prevention*, 2024, Vol. 208, p. 107812, doi:10.1016/j.aap.2024.107812.

[19] K. Zhang, D. Wu, Q. Liu, F. Dong, J. Liu, L. Jiang, and Y. Yuan, "Algorithm for drowsiness detection based on hybrid brain network parameter optimization," *Biomedical Signal Processing and Control*, 2024, Vol. 94, p. 106344, doi: 10.1016/j.bspc.2024.106344.

[20] X. Lin, Z. Huang, W. Ma, and W. Tang, "EEG-based driver drowsiness detection based on simulated driving environment," *Neurocomputing*, 2025, Vol. 616, p. 128961, doi:10.1016/j.neucom.2024.128961.

[21] X. Feng, S. Dai, and Z. Guo, "Pseudo-label-assisted subdomain adaptation network with coordinate attention for EEG-based driver drowsiness detection," *Biomedical Signal Processing and Control*, 2025, Vol. 101, p. 107132, doi:10.1016/j.bspc.2024.107132.

[22] F. Wang, M. Ma, R. Fu, and X. Zhang, "EEG-based detection of driving fatigue using a novel electrode," *Sensors and Actuators A: Physical*, 2024, Vol. 365, p. 114895, doi: 10.1016/j.sna.2023.114895.

[23] F. Wang, D. Chen, and X. Zhang, "Real-time Driving Fatigue Detection of ECG Signals Acquired Based on Novel Electrodes Using Wavelet Scattering Networks", *Measurement*, 2025, Vol. 243, p. 116438, doi:10.1016/j.measurement.2024.116438.

[24] Y. Liu, Z. Xiang, Z. Yan, J. Jin, L. Shu, L. Zhang, and X. Xu, "CEEMDAN fuzzy entropy based fatigue driving detection using single-channel EEG," *Biomedical Signal Processing and Control*, 2024,Vol. 95, Part A, p. 106460, doi:10.1016/j.bspc.2024.106460.

[25] I. Latreche, S. Slatnia, O. Kazar, and S. Harous, "An optimized deep hybrid learning for multi-channel EEG-based driver drowsiness detection," *Biomedical Signal Processing and Control*, 2025, Vol. 99, p. 106881, doi:10.1016/j.bspc.2024.106881.

[26] J. Chen, Y. Cui, H. Wang, E. He, and A. Alhudhaif, "Deep learning approach for detection of unfavorable driving state based on multiple phase synchronization between multi-channel EEG signals", *Information Sciences*,

2024, Vol. 658, p. 120070, doi:10.1016/j.ins.2023.120070.

[27] W. Yu, and Q. Huang, "A deep encoder-decoder network for anomaly detection in driving trajectory behavior under spatio-temporal context," *International Journal of Applied Earth Observation and Geoinformation*, 2022, Vol. 115, p. 103115, doi:10.1016/j.jag.2022.103115.

[28] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 International Conference on Engineering and Technology (ICET)*, Antalya, Turkey, 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

[29] L. Lin, S. Wang, J. Yang, and F. Wei, "A multi-aware graph convolutional network for driver drowsiness detection," *Knowledge-Based Systems*, 2024, Vol. 305, p. 112643, doi:10.1016/j.knosys.2024.112643.

[30] F. Wei, J. Yang, Y. Wang, L. Lin, and H. Zhang, "Prior knowledge-guided multi-information graph convolutional network for driver drowsiness detection", *Expert Systems with Applications*, 2025, Vol. 275, p. 127028, doi:10.1016/j.eswa.2025.127028.

[31] M. Elhenawy, M. Masoud, N. Haworth, K. Young, A. Rakotonirainy, R. Grzebieta, and A. Williamson, "Detection of driver distraction in the Australian naturalistic driving study videos using pre-trained models and transfer learning", *Transportation Research Part F: Traffic Psychology and Behaviour*, 2023, Vol. 97, pp. 31-43, doi:10.1016/j.trf.2023.06.016.

[32] B. Kanigoro, and B. Asdyo, "Facial Landmark and YOLOv5 Drowsiness Detection System," *Procedia Computer Science*, 2024, Vol. 245, pp. 548-554, doi:10.1016/j.procs.2024.10.281.

[33] Y. Ma, Z. Xie, S. Chen, F. Qiao, and Z. Li, "Real-time detection of abnormal driving behavior based on long short-term memory network and regression residuals", *Transportation Research Part C: Emerging Technologies*, 2023, Vol. 146, p. 103983, doi:10.1016/j.trc.2022.103983.

[34] N. Wang, T. Pu, Y. Zhang, Y. Liu, and Z. Zhang, "More appropriate DenseNetBL classifier for small sample tree species classification using UAV-based RGB imagery," *Heliyon*, 2023, Vol. 9, no. 10, p. e20467, doi:10.1016/j.heliyon.2023.e20467.

[35] L. Zhang, K. Yang, Y. Han, J. Li, W. Wei, H. Tan, P. Yu, K. Zhang, and X. Yang, "TSD-DETR: A lightweight real-time detection transformer of traffic sign detection for long-range perception of autonomous driving," *Engineering Applications of Artificial Intelligence*, 2025, Vol. 139, Part A, p. 109536, doi:10.1016/j.engappai.2024.109536.

[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, 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." Preprint, submitted June 3, 2021. Doi:10.48550/arXiv.2010.11929.

[37] T. Hugo, M. Cord, A. El-Nouby, J.Verbeek, and H. Jégou, "Three things everyone should know about vision transformers." In *Computer Vision–ECCV 2022*, edited by S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, vol. 13684, pp. 497-515. Cham: Springer Nature Switzerland, 2022, doi:10.1007/978-3-031-20053-3_29.

[38] P. Viola, and M. J. Jones, "Robust Real-Time Face Detection", *International Journal of Computer Vision*, 2004, vol. 57, pp. 137–154, doi:10.1023/B:VISI.0000013087.49260.fb.