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

ORIGINAL

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

Safe driving detection by Vision Transformer and convolutional networks comparison

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

Robinson Jiménez Moreno[1]
Anny Astrid Espitia Cubillos[1]
Javier Eduardo Martínez Baquero[2]

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

Received: 23 July 2025. Revised: 23 September 2025. Accepted: 20 December 2025. Published: 01 January 2026.
Corresponding Author: Anny Astrid Espitia Cubillos

How to cite: R. Jiménez Moreno, A. A. Espitia Cubillos, and J. E. Martínez Baquero, "Comparación de redes Vision Transformer y convolucionales para detección de conducción segura", *Revista Colombiana de Tecnologías de Avanzada (RCTA)*, vol. 1, no. 47, ene. 2026, art. no. 3824, doi: 10.24054/rcta.v1i47.3824

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

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.

INTRODUCTION

In recent years, ensuring safe driving [1] has become a fundamental concern due to the growing number of traffic accidents worldwide [2]. Advanced driver assistance systems (ADAS) and driver monitoring systems (DMS) [3] have emerged as promising solutions to improve 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 in the state of the art and associated with non-invasive systems such as artificial intelligence algorithms are convolutional neural networks (CNN), which have demonstrated remarkable success in image-based tasks [4] [5], and Vision Transformers (ViT), 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, traffic accidents claim approximately 1.35 million lives each year, and an even greater number of people suffer serious injuries [8]. One of the main contributors to these accidents is human error, often caused by drowsiness, distractions, or risky behaviors at the wheel. To address this challenge, advanced technologies have been developed to monitor driver behavior and detect signs of unsafe driving, with the aim of preventing accidents before they occur. Among these technologies, computer vision-based systems [9] have received significant attention because they can process visual data in real time and provide accurate assessments of the driver's state.

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 CNN and ViT[10][11]. CNN have been the cornerstone of image classification and object detection tasks for more than a decade, demonstrating impressive performance across a wide range of applications. Their hierarchical feature extraction leverages local connectivity and weight sharing, making them particularly suitable for image and video processing. However, CNN have certain limitations, such as their dependence on large labeled datasets and difficulties in capturing long-range dependencies within an image.

In contrast, ViT networks are a more recent 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, which allows the model to capture global context through self-attention mechanisms. This design lets ViTs overcome some of the shortcomings of CNNs, particularly when modeling long-range relationships and learning more holistic representations. Recent studies have shown that ViTs can outperform CNNs on several image recognition benchmarks, provided they are trained on sufficiently large datasets.

Research focused on safe driving systems largely focuses on autonomous technologies, especially object detection tasks such as pedestrian recognition and obstacle avoidance [13] [14]. However, research focused on driver-oriented safety systems remains highly relevant and continues to advance. These efforts address critical areas such as monitoring driver sobriety [15], evaluating driving behavior by detecting imbalances or erratic movements [16], and, most notably, detecting signs of drowsiness or drowsiness in the driver [17], where, for example, sleep detection requires clear identification of eye opening [18].

Although autonomous systems are designed to minimize human error, driver-centered approaches remain crucial for improving 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. Rather than replacing the driver, they work alongside the driver, offering support when needed and reducing risks on the road. Combined, autonomous technologies and driver-centered strategies create a more comprehensive and effective safety framework. This integrated approach not only addresses technical challenges but also takes the human element into account, 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 signal capture (EEG) [20] [21]. EEG has also been applied to detect driver fatigue [22], often combined with advanced feature extraction methods such as wavelet analysis [23] and fuzzy logic [24]. More recently, deep learning algorithms have demonstrated great performance in this area [25], successfully working both with EEG signals [26] and in identifying anomalies in driving behavior or trajectory [27]. These advances highlight the growing role of machine learning and signal processing in improving driver monitoring systems and road safety in general, with the limitation of needing to capture the patient's EEG signals.

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

The developments presented have demonstrated the advantages of deep networks in drowsiness detection; however, such works involve human intervention (EEG capture) or specificity of eye detection, which limits their applications in real time

and in real scenarios.

More recently, the state of the art uses deep learning models such as long short-term memory networks [33], and for image detection, Transformers networks are gaining strength [34], which are also beginning to be validated in autonomous driving, such as traffic sign detection [35]. However, their 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 CNN architecture and a pre-trained ViT model[36][37], to detect driving states classified as normal, distracted or drowsy. By evaluating the performance of these two approaches, the study contributes to the state of the art of driver monitoring systems, providing the inference response times of each network in real driving scenarios, under non-invasive systems not focused on the eye, which gives greater generality to the learning of the fatigue pattern.

The comparison aims to highlight the strengths and limitations of each model, especially with regard to their robustness and generalization capacity under various driving conditions. This analysis provides valuable information on which architecture may be most suitable for real-world applications, helping to inform the development of more reliable and efficient systems to improve driver safety and reduce traffic accidents.

This article is divided into four sections: the introduction, which presents the state of the art and the objective of this work; the methodology, which sets out the characteristics of the database and the architectures used; the analysis of results, which shows the performance and classification characteristics; and finally the conclusions.

METHODOLOGY

The proposed methodology, based on applied research, aims to establish a database under real 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. In turn, with the same database, transfer learning is used with the ViT architecture to obtain metrics of accuracy levels, inference time, F1 score and recall and, derived from the number of learning parameters, the size of the network, which affects memory usage in a real application. Table 1 illustrates the software and hardware characteristics used. Finally, the results obtained are presented alongside some of those reported for the 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 the CNN and ViT networks in identifying states relevant to safe driving, a user database was built across three different scenarios. The first scenario, called “normal”, represents the driver’s attention directed straight at the steering wheel and the road. The second scenario, “distraction”, captures the moments when the driver’s eyes wander, causing a loss of focus on the surroundings. Finally, the “sleep” state is characterized by the driver having their eyes closed or their head tilted down, which indicates drowsiness or microsleep episodes. This database makes it possible to evaluate how effectively each network can distinguish between these critical conditions in order to improve driving safety systems.

Fig. 1 shows part of the database used, in which the sleep, distraction and normal driving states appear with different users at 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 to validation and the remaining 10% were reserved for testing. This distribution ensures that the models are effectively trained, tuned and evaluated on different data. The database was built within a daytime lighting time range from 6 am to 6 pm, where each capture was made for a balanced distribution of each of the three established classes, ensuring that the same pose per user was not repeated in the distribution of the subsets. The diverse representation of users in the database helps to improve the generalization and robustness of the models when applied to real-world driving scenarios.

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 convolutional learning core (N) has the following notations: C convolution, B batch normalization, R for the Relu linear rectification unit, P for the pooling dimensionality reduction layer and, for the classification stage, FC as fully connected. The learning kernel per convolution core (N) consists 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 used 50% Dropout and linear activation functions (RELU).



Fig. 1. *Extract from the initial database*

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 N5 N6 N7	C-R-P C-R-P C-R-P C-B-R	3/48 4/96 4/96 [3 4/192]	2/1/1 2/0/1 2/0/1 0/1/2
FC	1024-2048-3		

For the case of the pre-trained ViT network, it is based on transfer learning under the model presented in [36] [37] [38]. This model has 143 layers, where the input image is managed based on 16 patches and for which fine adjustment is made in the network training by freezing the layers minus the attention layer and modifying the output layer to the target classes. As an activation function, the GELU and 10% Dropout scheme are used. A data augmentation technique based on rotation and reflection of the initial database is employed.

Each network is trained with the final parameters shown in Table 3. Note that the input volume is conditioned by the ViT network through transfer learning to a square image of 384 pixels per side. The images in the database are preprocessed during resizing to preserve the aspect ratio of the original image.

Table 3: Training parameters

Parameter	CNN	ViT
Input	384x384x3	384x384x3
Learning rate	0.00001	0.0001
Epochs	80	50
Mini-batch	12	4
Optimizer	ADAM	ADAM

RESULTS

After the training process, the accuracy graph initially indicates superior performance by the CNN, as shown in Fig. 2. The CNN reaches an accuracy of 91%, significantly surpassing the ViT, which only reaches an accuracy of 64.4%, as shown in Fig. 3. This notable difference reveals a possible advantage of the CNN in learning key features in the early phases of training. The results suggest that the CNN architecture is more suitable for capturing the relevant patterns required for this task, whereas the ViT may require further tuning or larger datasets to improve its performance.

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

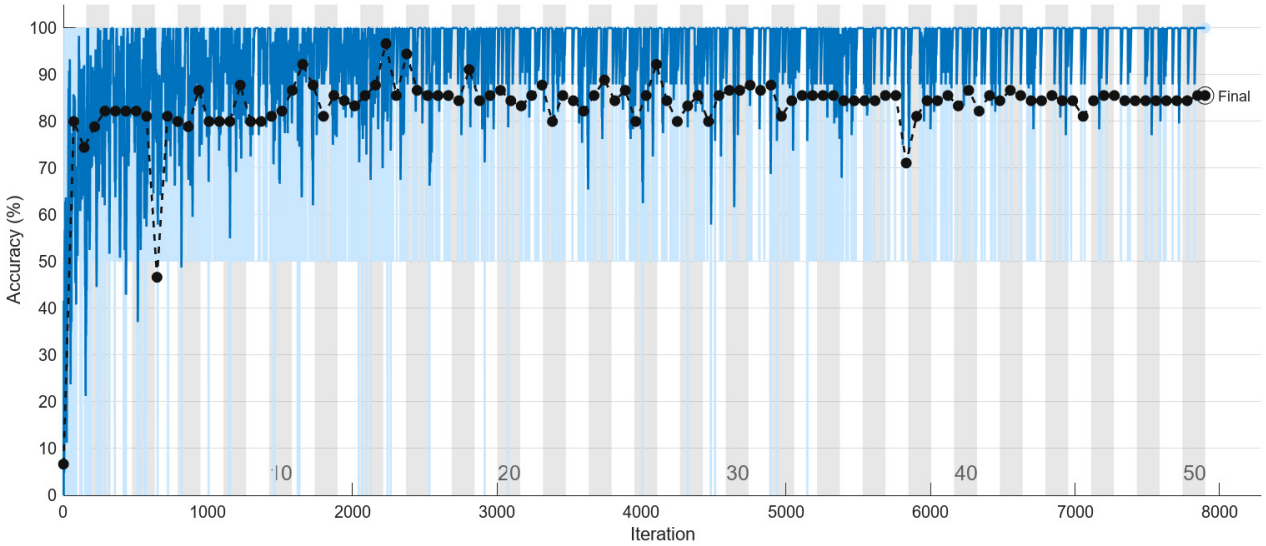

Fig. 2. *Initial training graph of the CNN*

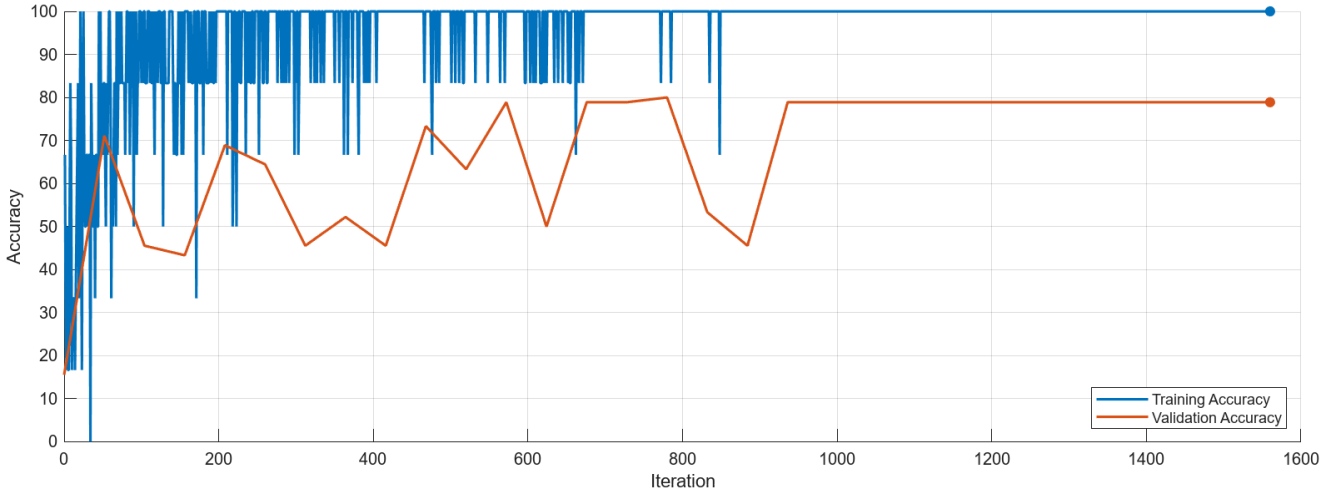

Fig. 3. *Initial training graph of the ViT*

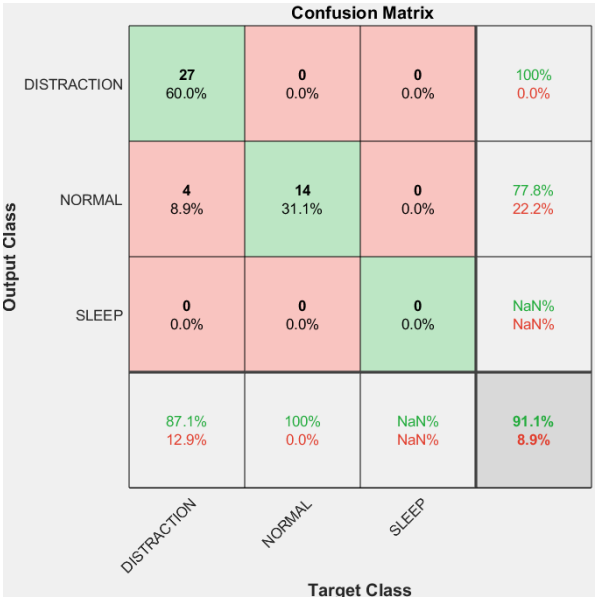

Fig. 4. *Initial confusion matrix of the CNN*

Table 4: Validation summary by class

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

Based on these results, it is determined to use a Haar classifier for facial recognition[26], applied to each initial image, thus generating a new set of images for training. Fig. 5 shows an excerpt from the updated database, in which the top row represents images of the sleep state, the middle row images of the distracted state and the bottom row images of normal driving. This approach improves the dataset by focusing on key facial features, which enables a more accurate classification of driver states. Next, the refined database is used to train driver state detection models, which improves the performance and robustness of the system.



Fig. 5. *Excerpt from the face database*

After retraining the networks using the same parameters listed in Table 3 and the face database, performance results were obtained with an accuracy of 98.9% for the CNN and 87% for the ViT.

Figures 6 and 7 show the performance results of each network. These results show the higher accuracy of the CNN and its effectiveness in detecting driver states, while the ViT, although slightly less accurate, still shows strong performance. Both networks are promising for driver monitoring applications, although the customized CNN proves more effective in this case.

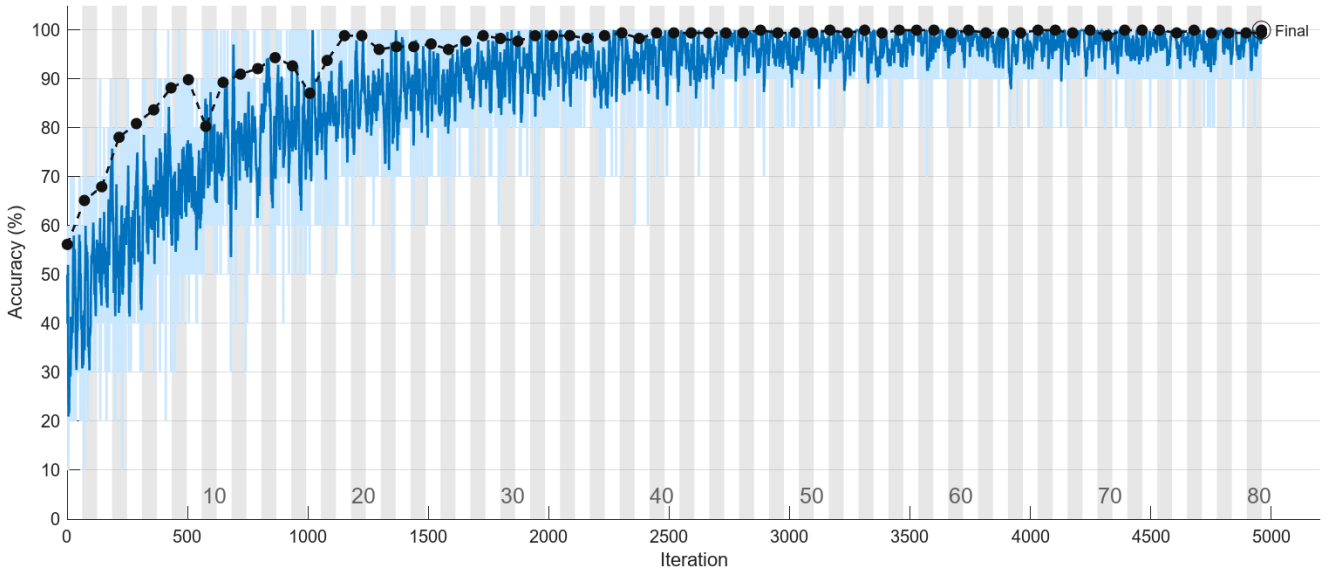

Fig. 6. *Training graph of the CNN with faces*

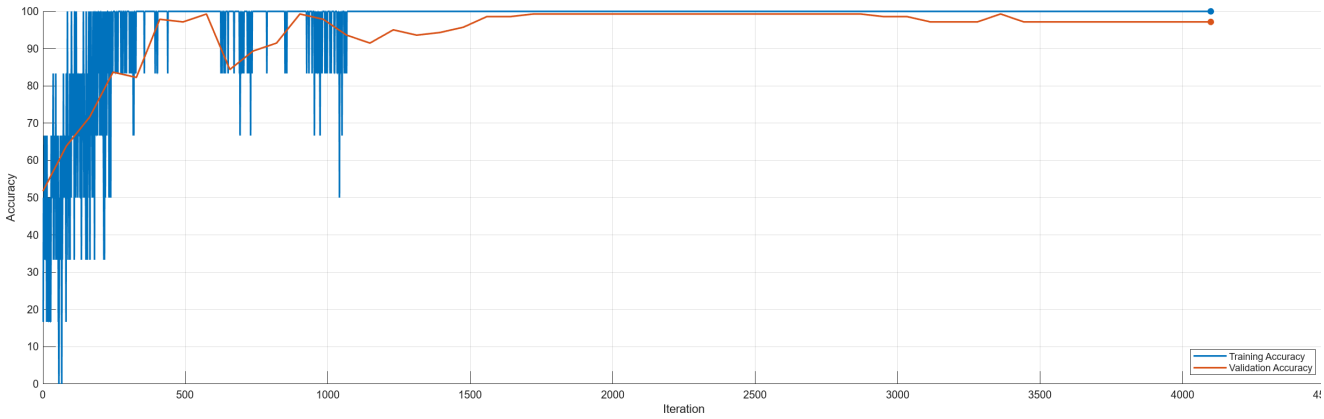

Fig. 7. *Training graph of the ViT with faces*

Fig. 8 shows the confusion matrix obtained for the CNN and Fig. 9 the confusion matrix obtained for the ViT. These results are tabulated in Table 5 and show the improvement in the result for each class for the CNN. Strong confusion is observed in the CNN in the distraction class, mainly with sleep. This behavior is also evident in the ViT but to a greater extent, and this class generates the main difference between the recognition of each type of network. In the case of safe driving and sleep detection, both networks identify a high percentage of these categories, with the CNN achieving 100% recognition, which is fundamental.

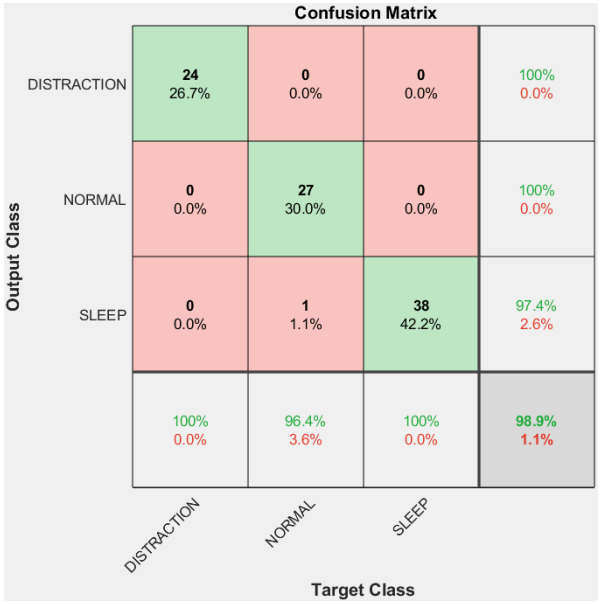

Fig. 8. *CNN confusion matrix*

Table 5: Summary of final validation by class

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

Table 6 graphically shows the results of both networks. The CNN shows a bias toward the distraction class and confusion between sleep and distraction. In contrast, the ViT model correctly identified each class without such confusion. In particular, an image showing a pronounced yawn, absent from the training dataset, was tested, and both networks systematically classified it in the distraction category. This example highlights the CNN's tendency to misclassify similar behaviors, whereas the ViT showed greater generalization. The agreement on the yawn image also underscores the challenges posed by unseen data and ambiguous facial expressions in classification tasks.

Extreme operating conditions of the algorithm were likewise analyzed, as can be seen in Fig. 10. Strong tilts or lateral turns of the face are delimited by the operation of the Haar classifier, for which the network's detection was conditioned on 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."

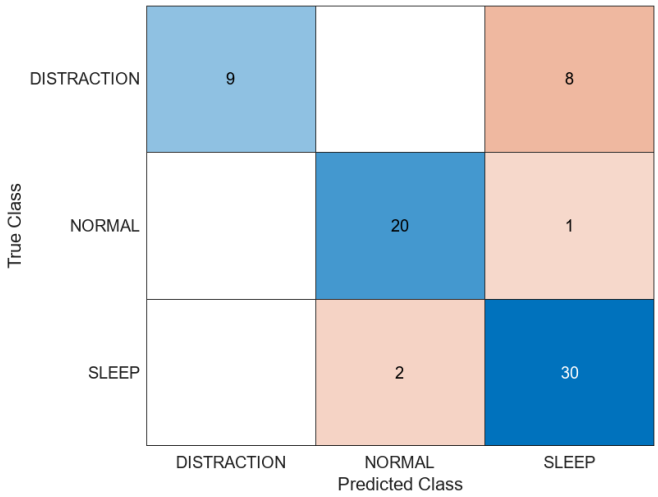

Fig. 9. *ViT confusion matrix*

The lower part of Fig. 10 shows some detection errors in the sleep class that are due to aspects such as eye occlusion (lower left), a fundamental aspect for identifying this state. Or, as Fig. 9 shows with respect to the confusion matrix, there are detection errors between the distraction class and the sleep class, where in Fig. 10 (lower right), when looking downward and the eye is partially visible without evidence of the sclera, the system detects it as sleep.

Table 6: *Graphical validation results*

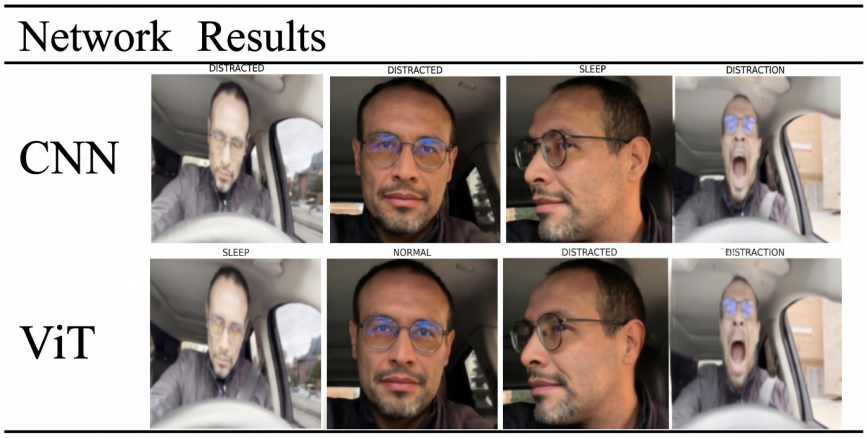



Fig. 10. Condiciones extremas de operación

The results show that the CNN, with only 30 layers, is almost five times faster than the ViT, which contains 143 layers. This significant difference highlights the CNN's advantage in terms of speed and computational efficiency, making it more suitable for real-time applications. In contrast, although the ViT network 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 for training, such as a GPU with at least 12GB of RAM.

A test was performed to evaluate the inference time of each network and to give an idea of its impact on a real-time system and the number of frames per second at which the detection system should operate. For the estimated calculation of the average inference time, a dataset of 25 images per class was used, and the results are summarized in Table 7, where it can be concluded that detection could occur at a maximum of 10 fps using the CNN network; the ViT would operate too slowly for a useful system.

For the quantitative analysis, 2000 results of the confidence level of the CNN detection are taken and tabulated in Table 8. The confidence distribution per class can be inferred to be very high for the normal and sleep states and, to a lesser degree, for the distracted state, which presents the highest variability at 31%. This means that the system is very reliable in the positive detections of each class, with a 75% percentile of confidences close to 1 in sleep detection, the most critical state of those analyzed by the network.

Table 7: Average inference times

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

Table 8: Final validation statistics per class

Class	Distraction	Normal	Sleep
Mean	0.722	0.927	0.923
Median	0.988	1	1
Standard deviation	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 from three users with different scenarios and durations were used. The driving state detection algorithm includes an alert graph to monitor the driver's state over time. Fig. 11 presents the driver's state throughout the observation window, showing three distinct levels determined by the result of the network classification. 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 does not detect the driver's face, the state value drops to zero, signaling the absence of detection. This graphical representation helps visualize the transitions between states and highlights the critical moments in which the driver's face is not recognized, providing essential information for safety interventions and ensuring that timely alerts are issued when risk levels increase.

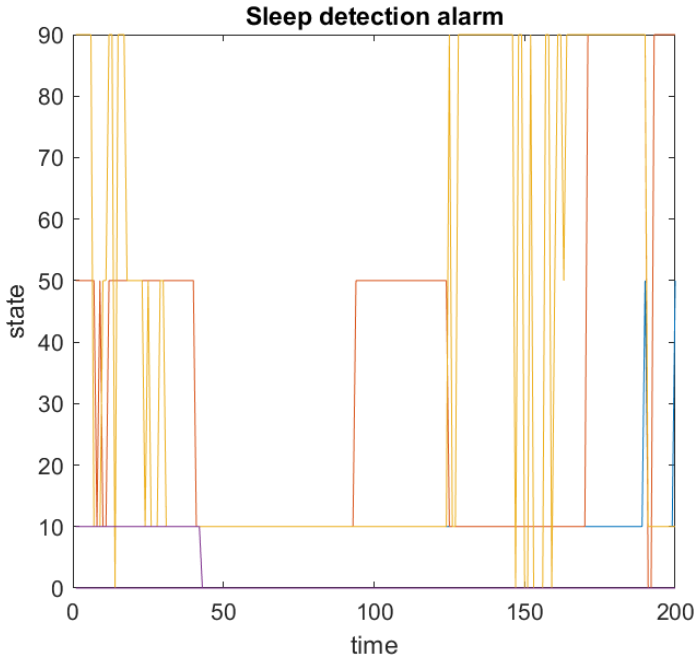

Fig. 11. *Sleep alert graph*

The applied algorithm yields dominance metrics for the driving task, in which the driver's fatigue state is determined by equations (1), (2), and (3).

$$I_s = \frac{Fsueño}{Ftotal} \tag{1}$$

$$I_d = \frac{Fdistracción}{Ftotal} \tag{2}$$

$$I_F = \frac{Fsueño + Fdistracción}{Ftotal} \tag{3}$$

Table 9 illustrates the dominance metrics in the real-scenario test videos. These results allow inferring aspects such as that driver three (video 3) drives in a fatigued state for a longer time, so if recovery breaks are established for this driver, the 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%

CONCLUSIONS

Regarding the results for accuracy, F1-score and inference times, it can be concluded that for the application of detection of the established safe driving states, a CNN performs better, presenting 11.8% higher classification accuracy than the ViT and with average inference times 5.14 times lower compared with the ViT.

The per-class statistical metrics of the CNN architecture show that detection of the most critical state, corresponding to sleep, is reliable, where the inference level is above 0.95 for 90% of the detections. Although detection of the distracted state can be improved, its confusion with the sleep class still generates driving alerts that favor the response of the proposed system.

As future work, evaluation based on public databases is proposed, 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-Transformers architectures.

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.

ACKNOWLEDGMENTS

The authors thank Universidad Militar Nueva Granada and Universidad de los Llanos, where they are tenured professors. Product derived from the research project entitled "Design of a human-robot interaction model using deep learning algorithms" INV-ING-3971, funded by the research vice-rectory of Universidad Militar Nueva Granada, 2024 term.

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Robinson Jiménez Moreno, Anny Astrid Espitia Cubillos, Javier Eduardo Martínez Baquero.

Writing–review and editing: Robinson Jiménez Moreno, Anny Astrid Espitia Cubillos, Javier Eduardo Martínez Baquero.