ISSN: 1692-7257 - Volumen 1 – Número 47 - 2026

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

Digital Object Identifier: 10.24054/rcta.v1i47.3824

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

Safe driving detection by Vision Transformer and convolutional networks comparison

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.*

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

Recibido: *23 julio 2025.* ***Aceptado:*** *20 diciembre 2025.* ***Publicado:*** *01 enero 2026.*

Cómo citar: *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, ene. 2026. Recuperado de https://ojs.unipamplona.edu.co/index.php/rcta/article/view/3824*

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.

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.

1. INTRODUCCIÓN

En los últimos años, garantizar una conducción segura [1] se ha convertido en una preocupación fundamental debido al creciente número de accidentes de tráfico en todo el mundo [2]. Los sistemas avanzados de asistencia al conductor (ADAS) y los sistemas de monitorización del conductor (DMS) [3] han surgido como soluciones prometedoras para mejorar la seguridad vial mediante la detección de signos de somnolencia, distracción o comportamiento de conducción inseguro. Con el rápido avance del aprendizaje profundo, las técnicas de visión por ordenador han desempeñado un papel central en la mejora del rendimiento de estos sistemas. Entre los enfoques más destacados, dentro del estado del arte y asociados a sistemas no invasivos como los algoritmos de inteligencia artificial, se encuentran las redes neuronales convolucionales (CNN), que han demostrado un éxito notable en tareas basadas en imágenes [4] [5], y las redes Transformers de Visión (ViT), una arquitectura más reciente que ha mostrado resultados sobresalientes en diversas aplicaciones de visión [6].

La seguridad vial se ha convertido en una de las preocupaciones mundiales más acuciantes de los últimos años [7]. Según la Organización Mundial de la Salud, los accidentes de tráfico cobran aproximadamente 1,35 millones de vidas cada año, y un número aún mayor de personas sufren lesiones graves [8]. Uno de los principales responsables de estos accidentes es el error humano, a menudo causado por la somnolencia, las distracciones o los comportamientos de riesgo al volante. Para hacer frente a este reto, se han desarrollado tecnologías avanzadas para supervisar el comportamiento del conductor y detectar señales de conducción insegura, con el objetivo de prevenir accidentes antes de que se produzcan. Entre estas tecnologías, los sistemas basados en visión por ordenador [9] han recibido una atención significativa debido a su capacidad para procesar datos visuales en tiempo real y proporcionar evaluaciones precisas del estado del conductor.

En el campo de la visión por ordenador, el aprendizaje profundo ha surgido como un enfoque transformador que permite a los modelos aprender patrones y características complejas directamente a partir de los datos basados en imágenes. Dos arquitecturas líderes en este campo son las CNN y ViT [10] [11]. Las CNN han sido la piedra angular de las tareas de clasificación de imágenes y detección de objetos durante más de una década, demostrando un rendimiento impresionante en una amplia gama de aplicaciones. Su extracción jerárquica de características aprovecha la conectividad local y el reparto de pesos, lo que las hace especialmente adecuadas para el procesamiento de imágenes y vídeos. Sin embargo, las CNN tienen ciertas limitaciones, como su dependencia de grandes conjuntos de datos etiquetados y las dificultades para captar las dependencias de largo alcance dentro de una imagen.

En contraste, las redes ViT representan un enfoque más reciente que aplica la arquitectura de Transformers, diseñada originalmente para tareas de procesamiento del lenguaje natural, a problemas de visión [12]. Las ViT dividen las imágenes en parches y las procesan como secuencias, lo que permite al modelo capturar el contexto global mediante mecanismos de autoatención. Este diseño permite a las ViT superar algunas de las deficiencias de las CNN, en particular a la hora de modelar

relaciones de largo alcance y aprender representaciones más holísticas. Estudios recientes han demostrado que los ViT pueden superar a las CNN en varias pruebas de referencia de reconocimiento de imágenes, siempre que se entrenen con conjuntos de datos suficientemente grandes.

Las investigaciones orientadas a los sistemas de conducción segura se centran en gran medida en las tecnologías autónomas, especialmente en tareas de detección de objetos como el reconocimiento de peatones y la evasión de obstáculos [13] [14]. Sin embargo, la investigación centrada en los sistemas de seguridad orientados al conductor sigue siendo muy relevante y continúa avanzando. Estos esfuerzos abordan áreas críticas como la supervisión de la sobriedad del conductor [15], la evaluación del comportamiento al volante mediante la detección de desequilibrios o movimientos erráticos [16] y, lo que es más notable, la detección de signos de somnolencia o sueño en el conductor [17], donde, por ejemplo, para la detección del sueño se requiere una clara identificación de la apertura del ojo [18].

Aunque los sistemas autónomos están diseñados para minimizar los errores humanos, los enfoques centrados en el conductor siguen siendo cruciales para mejorar la seguridad vial. Al supervisar continuamente el estado físico y cognitivo del conductor, estos sistemas pueden proporcionar alertas oportunas y ayudar a prevenir accidentes causados por la fatiga, la distracción u otros factores humanos. En lugar de sustituir al conductor, trabajan a su lado, ofreciéndole apoyo cuando lo necesita y reduciendo los riesgos en la vía. Cuando se combinan, las tecnologías autónomas y las estrategias centradas en el conductor crean un marco de seguridad más completo y eficaz. Este enfoque integrado no sólo aborda los retos técnicos, sino que también tiene en cuenta el elemento humano, contribuyendo en última instancia a crear entornos de conducción más seguros y fiables para todos los que circulan por la carretera.

Se están desarrollando nuevos modelos de aprendizaje para mejorar la detección del sueño del conductor, utilizando enfoques como las redes híbridas [19] y la captura de señales electroencefalográficas (EEG) [20] [21]. El EEG también se ha aplicado para detectar la fatiga del conductor [22], a menudo combinado con métodos avanzados de extracción de características como el análisis wavelet [23] y lógica difusa [24]. Más recientemente, los algoritmos de aprendizaje profundo han demostrado un gran rendimiento en este ámbito [25], trabajando con éxito tanto con señales de EEG [26] como en la identificación de anomalías en el comportamiento o la trayectoria de conducción [27]. Estos avances resaltan el creciente papel del aprendizaje automático y el procesamiento de señales en la mejora de los sistemas de supervisión de conductores y de la seguridad vial en general, con la limitante de requerir capturar las señales EEG del paciente.

Entre los principales algoritmos de aprendizaje profundo se encuentran las redes neuronales convolucionales CNN [28], que también han demostrado ser eficientes en la detección de conductores con sueño con base en arquitecturas ResNet [29] [30]. En este caso, existen modelos pre-entrenados por transferencia de aprendizaje [31], que utilizan arquitecturas CNN robustas como la red YOLO [32], basadas en la identificación ocular.

Los desarrollos expuestos han demostrado las ventajas de las redes profundas en detección de somnolencia, sin embargo, dichos trabajos implican intervención del humano (captura EEG) o especificidad de la detección ocular, lo cual delimita sus aplicaciones en tiempo y escenarios reales.

Más recientemente en el estado del arte se utilizan modelos de aprendizaje profundo como las redes de memoria a corto y largo plazo [33] y para la detección de imágenes están ganando fuerza las redes Transformers [34], que también se están empezando a validar en conducción autónoma como la detección de señales de tráfico [35]. Sin embargo, no es clara su ventaja frente a algoritmos CNN orientados a detección de sueño en un entorno de conducción segura al momento de revisión de la literatura.

En línea con las investigaciones expuestas y las ventajas de los algoritmos de aprendizaje profundo para conducción segura, este trabajo presenta una comparación entre una arquitectura de CNN y un modelo pre-entrenado ViT [36] [37], para detectar estados de conducción clasificados como normal, distraído o somnoliento. Al evaluar el rendimiento de estos dos enfoques, el estudio contribuye al estado del arte de los sistemas de monitorización de conductores, aportando los tiempos de respuesta en la inferencia de cada red en escenarios reales de conducción, bajo sistemas no invasivos ni centrados en el ojo, lo que da más generalidad al aprendizaje del patrón de cansancio.

La comparación pretende resaltar los puntos fuertes y las limitaciones de cada modelo, sobre todo en lo

que respecta a su solidez y capacidad de generalización en diversas condiciones de conducción. Este análisis proporciona información valiosa sobre qué arquitectura puede ser más adecuada para las aplicaciones del mundo real, ayudando a informar sobre el desarrollo de sistemas más fiables y eficientes para mejorar la seguridad del conductor y reducir los accidentes de tráfico.

Este artículo se divide en cuatro secciones, la introducción con una exposición del estado del arte y el objetivo de este trabajo. La metodología, donde se exponen las características de la base de datos y las arquitecturas utilizadas. El análisis de resultados, donde se muestran las características de rendimiento y clasificación, y por último las conclusiones.

2. METODOLOGÍA

La metodología propuesta, basada en investigación aplicada, se orienta a establecer una base de datos en condiciones reales de conducción con diferentes conductores. Dado que el estado del arte reporta resultados de arquitecturas basadas en CNN como las ResNet o YOLO, se propone una arquitectura propietaria CNN para comparación. A su vez con la misma base de datos se emplea transferencia de aprendizaje con la arquitectura ViT para obtener métricas de niveles de precisión, tiempo de inferencia, F1 score y recall y, derivado de la cantidad de parámetros de aprendizaje, el tamaño de la red que impacta el uso de memoria en una aplicación real. La tabla 1 ilustra las características de software y hardware empleadas. Finalmente se presentan los resultados conformados con alguno de los reportados para las arquitecturas ResNet y YOLO.

Tabla 1: *Configuración de software y hardware*

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

Para evaluar el rendimiento de las redes CNN y ViT en la identificación de estados relevantes para una conducción segura, se construyó una base de datos de usuarios en tres escenarios distintos. El primer escenario, denominado “normal”, representa la atención del conductor dirigida directamente hacia el volante y la vía. El segundo escenario, “distracción”, capta los momentos en los que los ojos del conductor se desvían, provocando una pérdida de concentración en el entorno. Por último, el estado de “sueño” se caracteriza porque el conductor tiene los ojos cerrados o la cabeza inclinada hacia abajo, lo que indica somnolencia o episodios de microsueño. Esta base de datos permite evaluar la eficacia con que cada red puede distinguir entre estas condiciones críticas para mejorar los sistemas de seguridad en la conducción.

La Fig. 1 muestra parte de la base de datos utilizada, en la que aparecen los estados de sueño, distracción y conducción normal con diferentes usuarios al volante. La base de datos completa consta de 670 imágenes de diez sujetos de prueba, que se dividieron en tres subconjuntos para el desarrollo del modelo: 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. Esta distribución garantiza que los modelos se entrenen, ajusten y evalúen eficazmente con datos distintos. La base de datos se construye en un rango horario de iluminación día en la franja de 6 am a 6 pm, donde cada captura se realiza para una distribución balanceada de cada una de las tres clases establecidas, asegurando que la misma pose por usuario no se repitiese en la distribución de los subconjuntos. La diversa representación de usuarios en la base de datos ayuda a mejorar la generalización y solidez de los modelos cuando se aplican a escenarios de conducción del mundo real.

Fig. 1. *Extracto de la base de datos inicial*

En el caso de la red convolucional (CNN), se utiliza la arquitectura ilustrada en la Tabla 2. Allí se utilizan las siguientes referencias para la etapa de extracción de características, donde la estructura de cada núcleo de aprendizaje convolucional (N) tiene las siguientes notaciones: C convolución, B Lote de normalización, R para la unidad de rectificación lineal Relu, P para la capa de reducción de dimensionalidad pooling y para la etapa de clasificación FC como completamente conectada. El kernel de aprendizaje por núcleo de convolución (N) se compone de filtros cuadrados de lado L y cantidad de filtros D bajo la relación L/D en la Tabla 2. A su vez se definen el tamaño de filtro de Maxpooling (M), el padding (P) y el paso o stride

(S), el paso de la operación de maxpooling se mantiene en 1. La etapa de clasificación empleo Dropout al 50% y funciones de activación lineal (RELU).

Tabla 2: *Arquitectura CNN*

Capa	Estructura	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		

Para el caso de la red ViT pre-entrenada, esta se basa en aprendizaje por transferencia bajo el modelo presentado en [36] [37] [38]. Este modelo cuenta con 143 capas, donde la imagen de entrada se maneja en base a 16 patches y para la cual se realiza un ajuste fino en el entrenamiento de la red congelando las capas menos la capa de atención y modificando la capa de salida a las clases objetivo. Como función de activación se emplea el esquema GELU y Dropout del 10%. Se emplea una técnica de aumento de datos basada en rotación y reflexión de la base de datos inicial.

Cada red se entrena con los parámetros finales que se muestran en la Tabla 3. Es importante destacar que el volumen de entrada se condiciona por la red ViT mediante transferencia de aprendizaje a una imagen cuadrada de 384 pixeles de lado. Se realiza un preprocesamiento en el redimensionamiento de las imágenes de la base de datos para mantener la relación de aspecto de la imagen original.

Tabla 3: *Parámetros de entrenamiento*

Parámetro	CNN	ViT
Entrada	384x384x3	384x384x3
Tasa de aprendizaje	0.00001	0.0001
Épocas	80	50
Mini lote	12	4
Optimizador	ADAM	ADAM

4. RESULTADOS

Tras el proceso de entrenamiento, el gráfico de precisión indica inicialmente un rendimiento superior por parte de la red CNN, como se muestra en la Fig. 2. La CNN alcanza una precisión del 91%, superando significativamente a la ViT, que sólo alcanza una precisión del 64,4%, como muestra la Fig. 3. Esta notable diferencia pone de manifiesto una posible ventaja de la CNN en el aprendizaje de características clave en las primeras fases del entrenamiento. Los resultados sugieren que la arquitectura de la CNN es más adecuada para captar los patrones relevantes necesarios para esta tarea, mientras que la red ViT puede requerir un mayor ajuste o conjuntos de datos más grandes para mejorar su rendimiento.

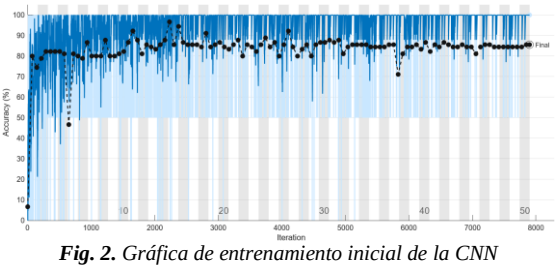

Fig. 2. *Gráfica de entrenamiento inicial de la CNN*

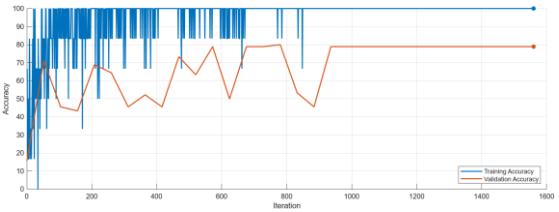

Fig. 3. *Gráfica de entrenamiento inicial de la ViT*

Sin embargo, al analizar la matriz de confusión de la red CNN (Fig. 4), es evidente que no discrimina bien entre las tres clases, eliminando la clase sueño, lo que se atribuye a que, dado el cambio de escala en la imagen, no es posible la identificación entre ojo cerrado y abierto. La Tabla 4 ilustra el desempeño obtenido por clase para lo cual la ViT exhibe un mejor comportamiento al identificar algo de cada clase.

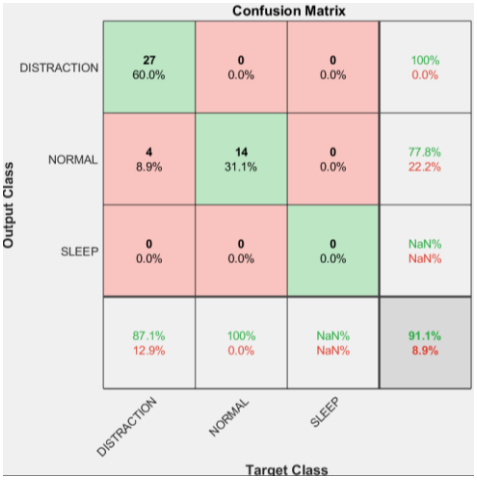

Fig. 4. *Matriz de confusión inicial de la red CNN*

Tabla 4: *Resumen validación por clase*

Clase	CNN(%)	ViT(%)
Distracción	60	31
Normal	31.1	23.4
Sueño	0	10

Derivado de estos resultados se determina emplear un clasificador Haar para el reconocimiento facial [26], aplicado a cada imagen inicial, generando así un nuevo conjunto de imágenes para el entrenamiento. La Fig. 5 muestra un extracto de la base de datos actualizada, en la que la fila superior representa imágenes del estado de sueño, la fila central imágenes del estado de distracción y la fila inferior imágenes del estado de conducción normal. Este enfoque mejora el conjunto de datos centrándose en los rasgos faciales clave, lo que permite una clasificación más precisa de los estados del conductor. A continuación, la base de datos refinada se utiliza para entrenar modelos de detección de los estados del conductor, lo que mejora el rendimiento y la solidez del sistema.

Fig. 5. *Extracto de la base de datos de rostros*

Tras reentrenar las redes utilizando los mismos parámetros que figuran en la Tabla 3 y la base de datos de rostros, se obtuvo unos resultados de rendimiento con una precisión del 98.9% para la red CNN y del 87% para la red ViT.

Las figuras 6 y 7 muestran los resultados de rendimiento de cada una de las redes. Estos resultados evidencian la mayor precisión de la CNN, mostrando su eficacia en la detección de los estados del conductor, mientras que la red ViT, aunque ligeramente menos precisa, sigue demostrando un gran rendimiento. Ambas redes resultan prometedoras en aplicaciones de supervisión de conductores, aunque la CNN personalizada demuestra ser más eficaz en este caso.

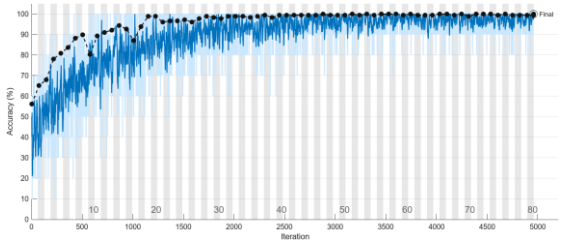

Fig. 6. *Gráfica de entrenamiento de la CNN con rostros*

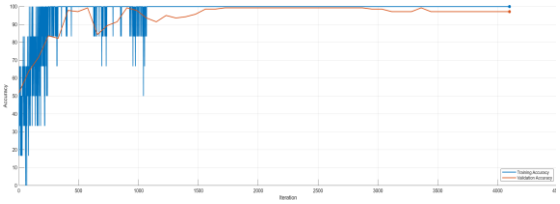

Fig. 7. *Gráfica de entrenamiento de la ViT con rostros*

La Fig. 8 muestra la matriz de confusión obtenida para la red CNN y la Fig. 9 la matriz de confusión obtenida para la red ViT. Estos resultados se tabulan en la Tabla 5 y evidencian la mejora en el resultado por cada clase para la CNN. Se observa una fuerte confusión en la CNN en la clase de distracción, principalmente con el sueño. Este comportamiento también es evidente en la ViT pero en mayor proporción, siendo esta clase la que genera la principal diferencia entre el reconocimiento de cada tipo de red. En el caso de la detección de conducción segura y sueño, ambas redes logran identificar un alto porcentaje de estas categorías, donde la CNN logra el 100% de reconocimiento, siendo fundamental.

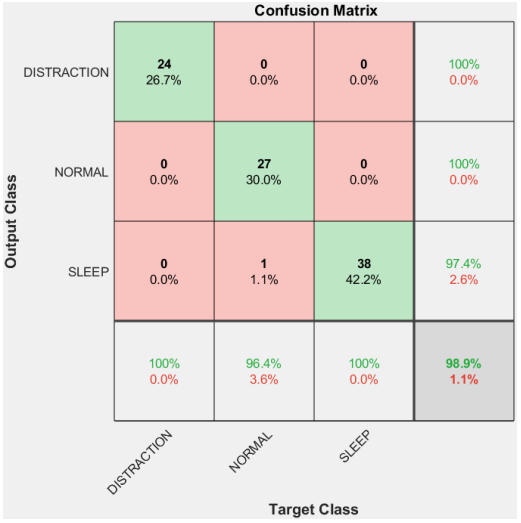

Fig. 8. *Matriz de confusión CNN*

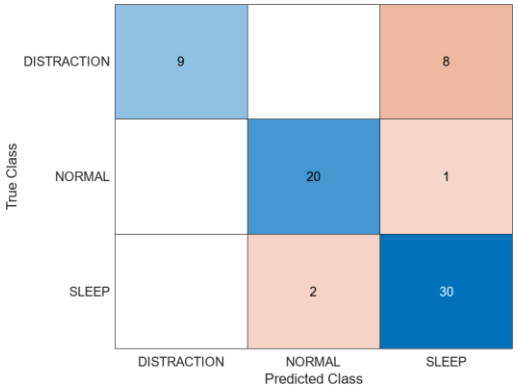

Fig. 9. *Matriz de confusión ViT*

Tabla 5: *Resumen validación final por clase*

Clase	CNN(%)	ViT(%)
Distracción	26.7	22.8
Normal	30	26.2
Sueño	42.2	38.1
Precisión	98.9	87.1
F1-score	98.9	82.2
Recall	98.8	80.6

La Tabla 6 permite evidenciar de forma gráfica los resultados de ambas redes. La red CNN presenta un sesgo de la clase de distracción y su confusión entre sueño y distracción. En cambio, el modelo ViT identificó correctamente cada clase sin dicha confusión. En particular, se probó una imagen que mostraba un bostezo pronunciado -ausente en el conjunto de datos de entrenamiento- y ambas redes la clasificaron sistemáticamente en la categoría de distracción. Este ejemplo pone de manifiesto la tendencia de la CNN a clasificar erróneamente comportamientos similares, mientras que el ViT demostró una mayor generalización. La coincidencia en la imagen del bostezo también subraya los retos que plantean los datos no vistos y las expresiones faciales ambiguas en las tareas de clasificación.

Tabla 6: *Resultados gráficos de validación*

Red	Resultados
CNN	
ViT	

De igual forma se analizaron condiciones extremas de operación del algoritmo que se pueden evidenciar en la Fig 10. Inclinaciones o giros laterales fuertes del rostro están delimitadas por la operación del clasificador Haar, para lo cual se condicionó la detección de la red al último valor reconocido a fin de evitar un falso estado de detección, tras 10 frames se genera una alarma como un cuarto estado denominado “no detección”.

Se pueden evidenciar en la parte inferior de la Fig. 10 algunos errores de detección en la clase de sueño que obedecen a aspectos como a una oclusión del ojo (parte izquierda inferior), aspecto fundamental para identificar este estado. O como lo evidencia la Fig. 9 respecto a la matriz de confusión, se presentan errores de detecciones en la clase distracción con la clase de sueño, donde en la Fig. 10 (parte derecha inferior), al estar mirando hacia abajo y verse parcialmente el ojo sin evidencia de la esclerótica, el sistema lo detecta como sueño.

Fig. 10. *Condiciones extremas de operación*

Los resultados muestran que la CNN, con sólo 30 capas, es casi cinco veces más rápida que la ViT, que contiene 143 capas. Esta significativa diferencia pone de manifiesto la ventaja de la CNN en términos de velocidad y eficiencia computacional, lo que la hace más adecuada para aplicaciones en tiempo real. En cambio, aunque la red ViT ofrece una mayor precisión y una mejor generalización, tiene el costo de unos tiempos de procesamiento más lentos debido a su arquitectura más profunda y a unos requisitos computacionales más complejos con relación al entrenamiento como una GPU de al menos 12GB de RAM.

Se realizó una prueba para evaluar el tiempo de inferencia de cada red y generar una idea de su impacto en un sistema en tiempo real y el número de frames (cuadros) por segundo al que debería operar el sistema de detección. Para el cálculo estimado del tiempo de la inferencia promedio se utilizó un conjunto de datos de 25 imágenes por clase, y los resultados se resumen en la Tabla 7, donde se puede concluir que la detección se podría dar máximo a 10 fps empleando la red CNN, la ViT operaria muy lento para un sistema útil.

Tabla 7: *Tiempos de inferencia promedio*

Red	Tiempo promedio de inferencia	FPS Max
CNN	0.102725	10
ViT	0.528490	2

Para un análisis cuantitativo se toman 2000 resultados del nivel de confidencia de la detección de la CNN que se tabulan en la Tabla 8. Se puede inferir que la distribución de confianza por clase es muy alta para los estados normal y sueño y en menor grado para el estado de distracción, que presenta la variabilidad más alta con un 31%. Lo que determina que el sistema es muy seguro en las detecciones positivas de cada clase, presentando un percentil del 75% con confianzas cercanas a 1 en la detección del sueño, el estado más crítico de los analizados por la red.

Tabla 8: *Estadísticas de validación final por clase*

Clase	Distracción	Normal	Sueño
Media	0.722	0.927	0.923
Mediana	0.988	1	1
Desviación estándar	0.315	0.224	0.2323
Varianza	0.099	0.0503	0.0539

Para validar la funcionalidad como una aplicación de detección de fatiga basada en las tres clases detectadas se emplearon 3 videos de tres usuarios con diferentes escenarios y duración. El algoritmo de detección del estado de conducción incluye un gráfico de alerta para controlar el estado del conductor a lo largo del tiempo. La Fig. 11 presenta el estado del conductor a lo largo de la ventana de observación, mostrando tres niveles distintos determinados por el resultado de la clasificación de la red. En concreto, un umbral de 10 corresponde a un estado de conducción normal, 50 indica un estado de distracción y 90 representa un estado de sueño. Cuando el sistema no detecta la cara del conductor, el valor del estado desciende a cero, señalando la ausencia de detección. Esta representación gráfica ayuda a visualizar las transiciones entre estados y destaca los momentos críticos en los que no se reconoce la cara del conductor, proporcionando información esencial para las intervenciones de seguridad y garantizando que se emitan alertas oportunas cuando aumenten los niveles de riesgo.

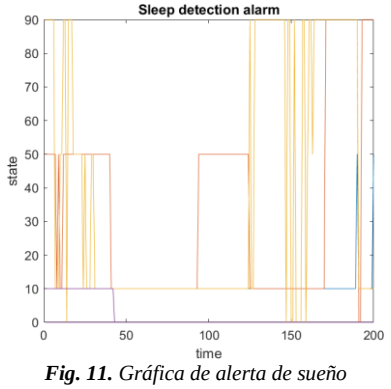

Fig. 11. *Gráfica de alerta de sueño*

Bajo el algoritmo aplicado se pueden obtener métricas de dominancia en la tarea de conducción, donde el estado de fatiga del conductor se determina mediante las ecuaciones (1), (2) y (3).

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

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

$$I_F = \frac{Fsueño+Fdistracción}{Ftotal} \quad (3)$$

La Tabla 9 ilustra las métricas de dominancia en los videos de prueba en escenario real, estos resultados permiten inferir aspectos como que el conductor tres (video 3) conduce en estado de cansancio mayor tiempo, por lo que si se establecen pausas de recuperación para este conductor los tiempos deberían ser menores.

Tabla 9: *Métricas de dominancia de fatiga*

Video	Sueño	Distracción	Fatiga
1	14.5 %	10%	27.4%
2	22.31%	9.18%	31.50%
3	16.8%	12.5%	44.3%

5. CONCLUSIONES

Se logra concluir en relación con los resultados de la precisión, F1-score y los tiempos de inferencia, que para la aplicación de detección de estados de conducción segura establecidos opera mejor una red CNN, la cual presenta un 11.8% mayor precisión en la clasificación que la ViT y con tiempos promedio de inferencia 5.14 veces menores en comparación con la ViT.

Las métricas estadísticas por clase de la arquitectura CNN evidencian que la detección del estado más crítico correspondiente a sueño es fiable, donde el nivel de inferencia esta por encima de 0,95 para el 90% de las detecciones. Si bien puede mejorar la detección del estado de distracción, su confusión con la clase de sueño sigue generando alertas de conducción que favorecen la respuesta del sistema propuesto.

Como trabajo futuro se establece realizar la evaluación en función a bases de datos públicas, así como la posible integración de una arquitectura hibrida CNN-LSTM para validar resultados basados en información temporal y explorar variantes más ligeras de arquitecturas Vision-Transformers.

AGRADECIMIENTOS

Los autores agradecen a la Universidad Militar Nueva Granada y a la Universidad de los Llanos,

donde son profesores titulares. Producto derivado del proyecto de investigación titulado “Diseño de un modelo de interacción humano-robot utilizando algoritmos de aprendizaje profundo" INV-ING-3971 financiado por la vicerrectoría de investigaciones de la Universidad Militar Nueva Granada, vigencia 2024.

REFERENCIAS

[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.