UNIVERSIDAD TÉCNICA DEL NORTE

FACULTAD DE INGENIERÍA EN CIENCIAS APLICADAS

CARRERA DE SOFTWARE

TRABAJO DE INTEGRACIÓN CURRICULAR PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO DE SOFTWARE

"APLICACIÓN DE TÉCNICAS DE INTELIGENCIA ARTIFICIAL EXPLICABLE EN MODELOS DE DEEP LEARNING PARA LA DETECCIÓN DE SOMNOLENCIA EN CONDUCTORES"

AUTOR:

Melany Masciel Sevilla Erazo

DIRECTOR:

Iván Danilo García Santillán

Ibarra - Ecuador

2026

IDENTIFICACIÓN DE LA OBRA

En cumplimiento del Art. 144 de la Ley de Educación Superior, hago la entrega del presente trabajo a la Universidad Técnica del Norte para que sea publicado en el Repositorio Digital Institucional, para lo cual pongo a disposición la siguiente información:

DATOS DE CONTACTO			
CÉDULA DE IDENTIDAD:	1004769327		
APELLIDOS Y NOMBRES:	SEVILLA ERAZO MELANY MASCIEL		
DIRECCIÓN:	DARIO EGAS Y PASAJE S/N		
EMAIL:	mmsevillae@utn.edu.ec		
TELÉFONO FIJO:		**TELF. MOVIL**	0986572578

DATOS DE LA OBRA	
TÍTULO:	APLICACIÓN DE TÉCNICAS DE INTELIGENCIA ARTIFICIAL EXPLICABLE EN MODELOS DE DEEP LEARNING PARA LA DETECCIÓN DE SOMNOLENCIA EN CONDUCTORES.
AUTOR (ES):	SEVILLA ERAZO MELANY MASCIEL
FECHA:	2026/03/27
SOLO PARA TRABAJOS DE INTEGRACIÓN CURRICULAR	
CARRERA/PROGRAMA:	■ **GRADO** □ **POSGRADO**
TITULO POR EL QUE OPTA:	INGENIERO EN SOFTWARE
DIRECTOR:	IVÁN DANILO GARCÍA SANTILLÁN
ASESOR	MARCO REMIGIO PUSDÁ CHULDE

CONSTANCIAS

La autora manifiesta que la obra objeto de la presente autorización es original y se la desarrolló, sin violar derechos de autor de terceros, por lo tanto, la obra es original y que es el titular de los derechos patrimoniales, por lo que asume la responsabilidad sobre el contenido de la misma y saldrá en defensa de la Universidad en caso de reclamación por parte de terceros.

Ibarra, a los 27 días, del mes de marzo de 2026.

EL AUTOR:

.....

Srta. Melany Masciel Sevilla Erazo

CERTIFICACIÓN DEL DIRECTOR DEL TRABAJO DE TITULACIÓN

Ibarra, 27 de marzo de 2026

PhD. Iván Danilo García Santillán

DIRECTOR DEL TRABAJO DE INTEGRACIÓN CURRICULAR

Por medio del presente yo PhD. Iván Danilo García Santillán, certifico que la Srta. Melany Masciel Sevilla Erazo portadora de la cedula de identidad número 1004769327, ha trabajado en el desarrollo del proyecto de grado **"Aplicación de técnicas de Inteligencia Artificial Explicable en modelos de Deep Learning para la detección de somnolencia en conductores"**, previo a la obtención del Título de Ingeniera en Software realizado con interés profesional y responsabilidad que certifico con honor de verdad.

Es todo en cuanto puedo certificar a la verdad

Atentamente

Firmado electrónicamente por:
IVAN DANILO GARCIA SANTILLAN
Validar únicamente con FirmaEC

……

PhD. Iván Danilo García Santillán

C.C.: 1002292603

APROBACIÓN DEL COMITÉ CALIFICADOR

El Comité Calificado del trabajo de Integración Curricular "APLICACIÓN DE TÉCNICAS DE INTELIGENCIA ARTIFICIAL EXPLICABLE EN MODELOS DE DEEP LEARNING PARA LA DETECCIÓN DE SOMNOLENCIA EN CONDUCTORES." elaborado por Melany Masciel Sevilla Erazo, previo a la obtención del título del Ingeniero en Software, aprueba el presente informe de investigación en nombre de la Universidad Técnica del Norte:

Firmado electrónicamente por:
IVAN DANILO GARCIA SANTILLAN
Validar únicamente con FirmaEC

……

PhD. Iván Danilo García Santillán

C.C.: 1002292603

Firmado electrónicamente por:
MARCO REMIGIO PUSDA CHULDE
Validar únicamente con FirmaEC

……

PhD. Marco Remigio Pusdá Chulde

C.C.: 0401200951

DEDICATORIA

A Dios, por estar presente en cada etapa de este camino y por darme la fortaleza necesaria en los momentos de mayor agotamiento. Gracias por la templanza para seguir cuando el cansancio quería más que la voluntad.

A mis padres, Patricio Sevilla y Jeaneth Erazo, por todo lo que han dado sin pedirlo y por estar presentes no solo en los logros, sino también en los tropiezos. Este trabajo lleva mucho de su esfuerzo y de su confianza en mí, y es el resultado de todo lo que sembraron desde el principio.

A mis hermanos, Erick y Matias, con quienes crecí y aprendí a moverme por el mundo. Su compañía ha sido una constante que doy por segura, pero que hoy reconozco y valoro con una gratitud que pocas veces expreso.

A mis abuelos, que partieron antes de poder ver este momento. Sus enseñanzas no se fueron con ellos, viven en la forma en que enfrento cada desafío y en los valores que guían lo que hago. Ojalá desde donde estén sepan que algo de lo que me dejaron está en este resultado.

A Kevin, Tatiana, Jordan y Melanie, por las risas dentro y fuera del aula, por las salidas que llegaron justo cuando más se necesitaban y por ese apoyo mutuo que no hace ruido pero que siempre se siente. Gracias por acompañarme de la manera en que solo los buenos amigos saben hacerlo.

Melany Masciel Sevilla Erazo

AGRADECIMIENTO

A toda mi familia, por ser el respaldo silencioso detrás de cada decisión y el lugar al que siempre supe que podía volver. Gracias por celebrar los logros como propios y por aguantar los momentos difíciles sin hacer demasiadas preguntas. Este camino fue más llevadero porque ustedes estuvieron en él.

Al PhD. Iván García, director de esta tesis, por guiar este proceso con paciencia y compromiso real, por compartir su conocimiento sin reservas y por ayudarme a encontrar el rumbo en los momentos en que no lo veía claro.

A la Universidad Técnica del Norte y a la carrera de Ingeniería en Software, por ser el espacio donde no solo adquirí conocimientos técnicos, sino también una forma de pensar y enfrentar los problemas que llevaré más allá del aula.

Y a todas las personas que fui conociendo en el trayecto, amigos que llegaron sin avisar y decidieron quedarse, gracias por los momentos compartidos, por las vivencias fuera de lo cotidiano y por haber hecho todo esto mucho más ligero y memorable.

Melany Masciel Sevilla Erazo

RESUMEN

La fatiga al conducir constituye un factor de riesgo crítico para la seguridad vial, no obstante, la naturaleza de "caja negra" de los modelos de Deep Learning empleados para su detección genera desconfianza y limita su validación técnica. Esta investigación tuvo como objetivo general aplicar técnicas de Inteligencia Artificial Explicable (XAI) a modelos de aprendizaje profundo para la detección de somnolencia, mejorando la transparencia y la comprensión de sus predicciones. La metodología se fundamentó en el estándar CRISP-DM, evaluando un modelo base mediante Grad-CAM, SHAP, LIME y Score-CAM para identificar sesgos contextuales; a partir de estos hallazgos, se implementó una estrategia de preprocesamiento optimizada con enmascaramiento dinámico y desenfoque gaussiano. Los resultados evidenciaron que el modelo optimizado alcanzó una exactitud del 96.16%, logrando una reducción del 51% en falsos negativos y un incremento de hasta 44 veces en la precisión de la localización anatómica de las activaciones. Se concluye que el uso de XAI permite mitigar el fenómeno de "Clever Hans" al alinear el enfoque del modelo con indicadores fisiológicos validados como el estándar PERCLOS, además de fomentar una calibración epistémica que permite al sistema expresar incertidumbre ante datos ambiguos, fortaleciendo así la seguridad operativa.

Palabras clave: Inteligencia Artificial Explicable, Deep Learning, Somnolencia del conductor, Grad-CAM, PERCLOS, Visión por computadora.

ABSTRACT

Driver fatigue is a critical road safety risk factor; however, the "black-box" nature of Deep Learning models used for its detection generates mistrust and limits technical validation. This research aimed to apply Explainable Artificial Intelligence (XAI) techniques to deep learning models for drowsiness detection, enhancing transparency and understanding of their predictions. The methodology was based on the CRISP-DM standard, evaluating a base model through Grad-CAM, SHAP, LIME, and Score-CAM to identify contextual biases; based on these findings, an optimized preprocessing strategy with dynamic masking and Gaussian blur was implemented. Results showed that the optimized model achieved 96.16% accuracy, achieving a 51% reduction in false negatives and up to a 44-fold increase in the accuracy of anatomical localization of activations. It is concluded that the use of XAI allows mitigating the "Clever Hans" phenomenon by aligning the model's focus with validated physiological indicators such as the PERCLOS standard, in addition to fostering an epistemic calibration that allows the system to express uncertainty in the face of ambiguous data, thus strengthening operational safety.

Keywords: Explainable Artificial Intelligence (XAI), Deep Learning, Driver drowsiness detection, Grad-CAM, PERCLOS, Computer Vision.

LISTA DE SIGLAS

XAI. Inteligencia Artificial Explicable (Explainable Artificial Intelligence).

IA. Inteligencia Artificial.

CNN. Redes Neuronales Convolucionales (Convolutional Neural Networks).

Grad-CAM. Mapeo de Activación de Clases pesado por Gradiente (Gradient-weighted Class Activation Mapping).

SHAP. Explicaciones Aditivas de Shapley (SHapley Additive exPlanations).

LIME. Explicaciones Locales Interpretables Independientes del Modelo (Local Interpretable Model-agnostic Explanations).

Score-CAM. Mapeo de Activación de Clases basado en Puntuación.

PERCLOS. Porcentaje de Cierre Ocular (Percentage of Eye Closure).

CRISP-DM. Metodología de Proceso Estándar de la Industria para la Minería de Datos.

OMS. Organización Mundial de la Salud.

LSTM. Redes de Memoria a Largo Plazo (Long Short-Term Memory).

GRU. Unidades Recurrentes de Compuerta (Gated Recurrent Units).

CI/CD. Integración Continua y Despliegue Continuo.

FPS. Cuadros por segundo (Frames Per Second).

ROC/AUC. Característica Operativa del Receptor / Área Bajo la Curva.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

CAPÍTULO I

INTRODUCCIÓN

1.1 Problema de investigación

Según un informe reciente de la OMS [1], los accidentes de tránsito siguen siendo una de las principales causas de fallecimiento a nivel mundial, con 1,19 millones de víctimas en 2021. La fatiga y la somnolencia son factores de alto riesgo para la seguridad vial. Los sistemas de detección de somnolencia basados en visión por computadora ofrecen un monitoreo continuo y no invasivo [2]. Sin embargo, los modelos de Deep Learning, aunque precisos en la clasificación de imágenes [3], funcionan como "cajas negras", lo que dificulta la comprensión de las decisiones que toman. Esta falta de transparencia genera desconfianza entre los usuarios, obstaculiza la validación técnica de las decisiones, impide el refinamiento de los sistemas y aumenta el riesgo en entornos críticos donde sus fallos pueden resultar fatales. Estudios recientes han demostrado que las técnicas de Explicabilidad en Inteligencia Artificial (Explainable Artificial Intelligence, XAI), aplicadas en modelos de Deep Learning para la detección de somnolencia en conductores, permiten generar explicaciones visuales sobre el funcionamiento de dichos modelos. Esto permite visualizar las regiones de una imagen que influyen en la toma de decisiones, lo cual es fundamental en aplicaciones críticas como la seguridad vial y la medicina.

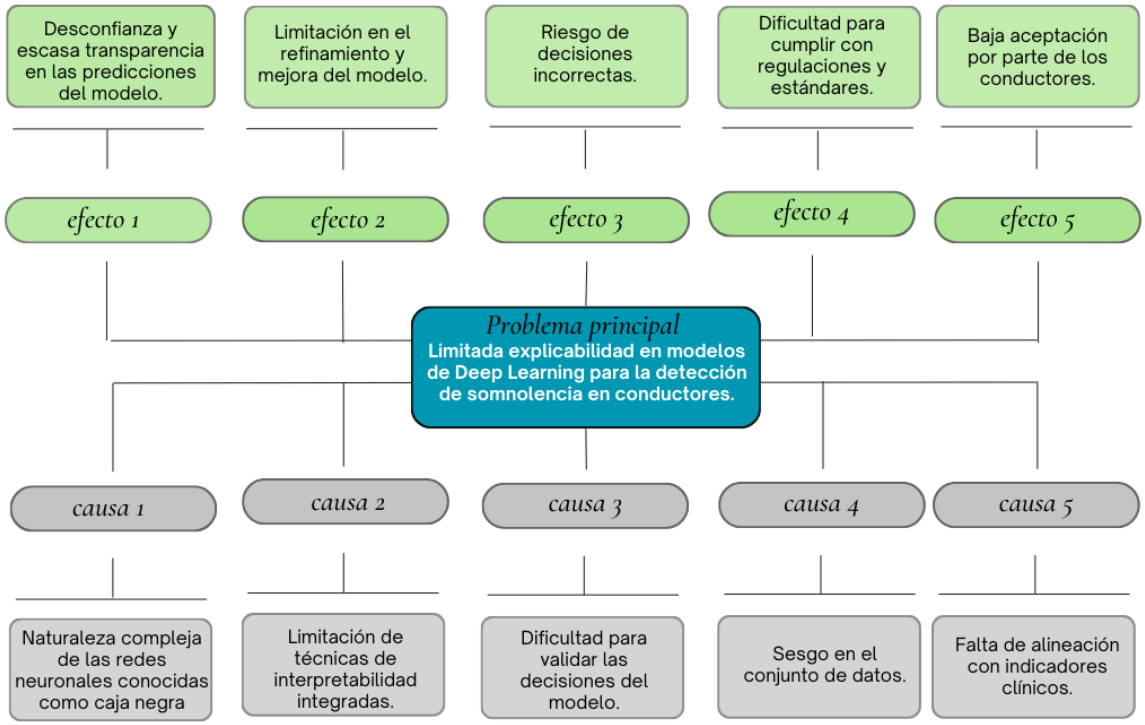

Figura 1. Árbol de problemas

1.2 Objetivos

1.2.1 Objetivo General

Aplicar técnicas de XAI a modelos de Deep Learning para la detección de somnolencia en conductores, mejorando la comprensión y explicación de sus predicciones de manera clara y transparente.

1.2.2 Objetivos Específicos

- Redactar un marco teórico sobre las técnicas de XAI utilizadas en modelos de Deep Learning para detección de somnolencia en conductores, con énfasis en SHAP, LIME, y Grad-CAM.
- Aplicar técnicas de explicabilidad XAI a un modelo preentrenado de somnolencia para comprender la coherencia y confiabilidad mediante métricas de interpretabilidad.

- Reentrenar el modelo de somnolencia incorporando los hallazgos obtenidos de la explicabilidad y evaluar el rendimiento del modelo utilizando métricas de IA.

1.3 Alcance

Este trabajo evaluará y mejorará la explicabilidad de un modelo de Deep Learning basado en redes neuronales convolucionales para la detección de somnolencia en conductores, con el objetivo de hacer más comprensibles sus predicciones. La explicabilidad en inteligencia artificial (XAI) permite que los modelos, que suelen ser complejos debido a sus múltiples capas y millones de parámetros, sean entendidos por los humanos a través de métodos que aclaran sus decisiones, ya que explican qué partes de la entrada (imagen) influyen en las decisiones del modelo [4].

Se aplicarán las técnicas de XAI mencionadas (SHAP, LIME, Grad-CAM) para realizar un análisis de las regiones relevantes de la imagen (por ejemplo, rostro y ojos del conductor) que influyen principalmente en la predicción de somnolencia. Se utilizará Grad-CAM [5] para generar mapas de calor visuales que muestran las áreas más influyentes de las imágenes, mientras que LIME [6] proporcionará explicaciones locales mediante perturbaciones de las características de entrada. Finalmente, se empleará SHAP [7] para medir la contribución de cada característica en la predicción del modelo. También, se explorará Score-CAM [8] debido a que parecería ser más robusto porque no depende de gradientes como Grad-CAM.

Con base en los hallazgos obtenidos, se ajustará el modelo para mejorar su desempeño y transparencia, evaluando posteriormente los resultados con las mismas métricas.

Se llevarán a cabo evaluaciones cuantitativas para medir el rendimiento del modelo antes y después de aplicar las técnicas de XAI. Las métricas de evaluación incluirán precisión, recall, F1-score y matrices de confusión [9] Además, se realizará un análisis cualitativo para determinar si las explicaciones generadas son consistentes y confiables en diversos escenarios.

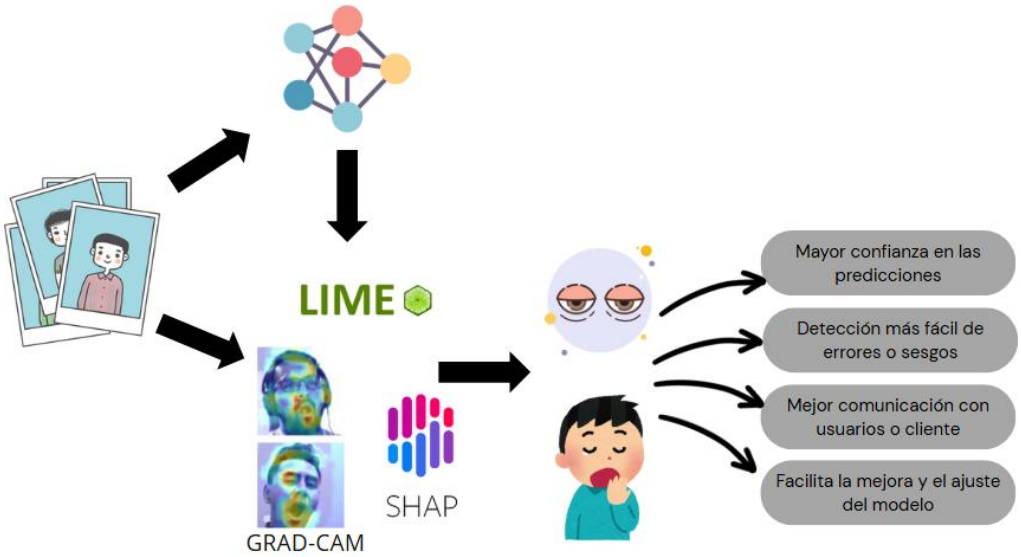

Figura 2.Diagrama propuesto para el proyecto.

1.4 Metodología

Este trabajo de investigación es de tipo aplicada y, para su desarrollo, se implementará la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), adaptada al enfoque de XAI [10]. Las etapas del proceso son las siguientes:

- **Comprensión del negocio:** Establecimiento de los objetivos del proyecto e identificación de las necesidades de modelos explicables para la detección de somnolencia.
- **Comprensión de los datos:** Para alcanzar el primer objetivo, se llevará a cabo una revisión de literatura sobre las técnicas XAI aplicadas a modelos de Deep Learning, consultando bases de datos científicas y recursos académicos para construir el marco teórico.
- **Preparación de los datos:** Configuración del entorno computacional (TensorFlow/PyTorch) y organización del conjunto de datos de conductores.
- **Modelado:** En relación con el segundo objetivo, se implementarán las técnicas XAI sobre el modelo preentrenado. Posteriormente, se realizará una evaluación tanto cuantitativa como cualitativa del modelo.
- **Evaluación:** Para alcanzar el tercer objetivo, se llevará a cabo un reentrenamiento del modelo basado en los hallazgos obtenidos en la evaluación de las técnicas, ajustando los hiperparámetros según las características más relevantes y descartando las menos significativas.

- **Despliegue:** Evaluación comparativa entre modelo original y reentrenado mediante métricas como precisión, recall, F1-score y UC-ROC. Además, se elaborarán recomendaciones basadas en evidencias para mejoras futuras.

1.4 Justificación

Este proyecto se alinea con el Objetivo de Desarrollo Sostenible (ODS) número 3, que tiene como propósito reducir las muertes y lesiones causadas por accidentes de tráfico [11]. Además, se vincula con la política 9.2 del Plan de Creación de Oportunidades 2021-2025 [12], la cual refuerza la seguridad de los sistemas de transporte terrestre y aéreo. Este enfoque tiene como objetivo mejorar la seguridad mediante el uso de tecnologías explicables, reduciendo así el riesgo de accidentes.

Beneficiarios

Conductores, profesionales del área de seguridad vial y empresas tecnológicas.

Justificación Académica

Contribuye al campo de la inteligencia artificial en visión por computadora, evaluando técnicas de XAI, lo que permite a los investigadores y desarrolladores identificar las fortalezas y limitaciones de los modelos para aplicaciones críticas.

Justificación económica

La incorporación de estos modelos optimizará recursos al reducir costos operativos asociados con la validación manual y los errores humanos, como malentendidos en predicciones que generan pérdidas económicas por decisiones erróneas o desconfianza. Además, una mejor transparencia generará mayor confianza y adopción del modelo en el sector productivo, mejorando la rentabilidad y sostenibilidad.

CAPÍTULO II

2. MARCO TEÓRICO

2.1 Inteligencia Artificial Explicable (XAI)

2.1.1 Definición y conceptos fundamentales

La Inteligencia Artificial (IA) es un campo en constante evolución que aún carece de una definición única y universalmente aceptada [13]. Sin embargo, la OCDE define los sistemas de IA como aquellos operados por máquinas que, con objetivos definidos por humanos, pueden realizar predicciones, recomendaciones o decisiones que impactan tanto en entornos reales como virtuales [13]. Dentro de este ecosistema, el Aprendizaje Automático (Machine Learning) y el Aprendizaje Profundo (Deep Learning) son subcampos que han impulsado grandes avances en tareas, como el procesamiento del lenguaje natural y la visión por computadora [14].

La Inteligencia Artificial Explicable (XAI) surge como una disciplina para abordar la complejidad y opacidad de los sistemas de IA modernos, esta se define como el conjunto de procesos y métodos que ayudan a los usuarios humanos a comprender y confiar en los resultados generados por algoritmos de IA [15]. Sin embargo, los conceptos claves de XAI se basan en tres principios interrelacionados que, por la falta de un acuerdo generalizado, a menudo se usan como si fueran lo mismo [15] :

- **Transparencia:** Se cumple cuando los procesos que calculan los parámetros del modelo y generan los resultados pueden describirse y justificarse [15]. Esto significa poder rastrear el origen de los datos y comprender el proceso de toma de decisión. Además, implica comunicar las capacidades y limitaciones del sistema a los usuarios [16].
- **Interpretabilidad:** Es la capacidad de comprender cómo el modelo toma decisiones de forma comprensible para los humanos. Generalmente, se relaciona con explicar las causas y efectos de sus resultados [15]. Existe una diferencia importante aquí: la IA interpretable utiliza modelos que ya son comprensibles por diseño, mientras que la IA explicable aplica

técnicas para entender modelos más complejos u opacos [16].

- **Explicabilidad:** Esta característica permite entender por qué una observación específica produjo un resultado particular, es decir, implica conocer la lógica interna del algoritmo, su diseño, entrenamiento y los pasos seguidos para llegar a un resultado [15].

Estos tres principios constituyen la base conceptual sobre la cual se desarrollan las técnicas y métodos de XAI que se describirán en las siguientes secciones.

2.1.2 Importancia de la explicabilidad en modelos de Deep Learning: El desafío de la caja negra

Los modelos de Deep Learning, aunque tienen un gran rendimiento, a menudo funcionan como "cajas negras" [15], [17], [18], lo que significa que se conocen sus resultados, pero el proceso de razonamiento interno es difícil de comprender para los humanos [14], [15]. A diferencia de los sistemas de IA tradicionales que sigue reglas claras, las redes neuronales profundas aprenden patrones complejos [17], creando una necesidad de explicabilidad, especialmente en aplicaciones de alto riesgo como la detección de somnolencia.

Esta necesidad surge por múltiples razones interconectadas. Primero, la adopción exitosa de cualquier tecnología depende de la confianza del usuario, y cuando estos no pueden comprender como un modelo toma decisiones, naturalmente se resisten a utilizarlo, limitando su impacto práctico [18]. De igual manera, en el ámbito regulatorio, el GDPR europeo establece el "derecho a una explicación" para decisiones automatizada, mientras que la Ley de IA de la UE requiere transparencia y explicabilidad para sistemas de alto riesgo [15], haciendo que las decisiones sean auditables y responsables [14].

Desde un enfoque técnico, la explicabilidad facilita la mejora continua de los modelos, ya que al entender qué características influyen en las predicciones, los desarrolladores pueden identificar sesgos, corregir errores y anticipar fallos [13], [15]. Esto también previene el mal uso y protege la reputación, la falta de transparencia puede llevar a decisiones discriminatorias o aplicaciones inadecuadas [15]. Finalmente, en contextos críticos como el diagnóstico médico o la seguridad vial, la explicabilidad es vital para proteger a las personas, permitiendo la intervención humana oportuna y asegurando la rendición de cuentas [14], [15], [17].

Para abordar el desafío, se emplean dos estrategias principales. La primera consiste en utilizar Algoritmos inherentemente interpretables o de “caja blanca”, que fueron diseñados para ser comprensibles por su propia naturaleza como árboles de decisión o regresiones lineales [15]. La segunda estrategia implica aplicar técnicas de interpretabilidad post-hoc después de entrenar modelos opacos, como LIME y SHAP [15], que permiten generar explicaciones sin modificar la arquitectura original del modelo.

2.1.3 Tipos de explicabilidad: global vs local

La interpretabilidad de los modelos se puede clasificar en dos categorías principales según el alcance de la explicación que proporcionan [16]:

Explicabilidad Global

Tiene como objetivo comprender la lógica general de decisión del modelo completo, proporcionando una visión amplia de cómo las características influyen en las predicciones a gran escala. Para esto se utilizan técnicas como árboles de interpretación y herramientas como los *Partial Dependence Plots* (PDP) e *Individual Conditional Expectation* (ICE), que revelan las reglas de decisión implícitas. Su principal ventaja radica en la capacidad de analizar el impacto general de los parámetros sin asumir relaciones específicas entre características y con poco esfuerzo computacional, aunque presenta desafíos en modelos muy complejos, ya que no existe un consenso sobre cómo medir ese impacto global [19].

Explicabilidad Local

Se centra en entender decisiones específicas para casos particulares, generando explicaciones individuales mediante la perturbación de entradas y el entrenamiento de modelos sustitutos interpretables que imitan el comportamiento del modelo original en el entorno cercano a cada predicción. Técnicas como LIME y SHAP son ejemplos comunes, que ofrecen la ventaja de conocer detalladamente cómo cada característica influyó en una predicción individual, lo cual es fundamental para generar confianza y rendición de cuentas. Sin embargo, este método asume independencia entre características y relaciones lineales, lo que puede generar explicaciones engañosas cuando existen interacciones complejas, además de requerir un alto costo computacional al generar explicaciones para cada caso nuevo [19].

2.1.4 Técnicas y métodos en XAI

Dentro de este campo se han desarrollado diversas técnicas para abordar el problema del bajo nivel de explicabilidad en los modelos. Estas técnicas pueden clasificarse según múltiples criterios complementarios que permiten seleccionar el método más apropiado según las necesidades del proyecto[14], [20], [21].

Clasificación por momento de aplicación:

- **Métodos Ante-hoc (Intrínsecos):** Se utilizan durante el desarrollo del sistema para construir modelos transparentes por naturaleza, como árboles de decisión o regresiones lineales [14]. Aunque son fáciles de entender, suelen tener menor capacidad predictiva que modelos más complejos [21].

- **Métodos Post-hoc:** Se aplican después de entrenar los modelos para explicar sus predicciones. Son comunes por su flexibilidad para trabajar con modelos pre-entrenados sin necesidad de modificar su arquitectura [14].

Dentro de los métodos post-hoc, se distinguen dos enfoques según su alcance:

Clasificación por alcance del modelo:

- **Técnicas Agnósticas del Modelo:** Funcionan con cualquier modelo, sin importar su arquitectura interna. Son versátiles y se centran en la relación entre entradas y salidas. Ejemplos representativos son LIME y SHAP [14], [20].

- **Técnicas Específicas del Modelo:** Diseñadas para ciertos tipos de modelos, aprovechando su estructura interna. Pueden no ser aplicables a otras arquitecturas. Ejemplos incluyen Grad-CAM para CNNs y las puntuaciones de atención para Transformers [14] .

Clasificación por metodología de análisis:

- **Basadas en Perturbación:** Modifican los datos de entrada para determinar qué tan importante es cada característica en la predicción, observando como cambia la salida del modelo [14].

- **Basadas en Gradiente:** Calculan los gradientes de la predicción con respecto a

datos de entrada, reflejando la sensibilidad matemática de la salida a cambios en cada característica [14].

Tabla 1.Taxonomía consolidada de técnicas XAI

Criterio de Clasificación	Categorías	Ejemplos Representativos	Ventajas Principales
Momento de aplicación	Ante-hoc	Árboles de decisión, Regresión lineal	Transparencia inherente, no requieren procesamiento adicional
	Post-hoc	LIME, SHAP, Grad-CAM	Aplicables a modelos complejos ya entrenados
Alcance del modelo	Agnósticas	SHAP, LIME, PDP	Versatilidad, aplicables a cualquier arquitectura
	Específicas	Grad-CAM, Score-CAM, Attention Maps	Mayor precisión al explotar estructura interna
Metodología	Perturbación	LIME, Occlusion Maps	Interpretación intuitiva, simulan razonamiento humano
	Gradiente	Grad-CAM, Saliency Maps	Computacionalmente eficientes, precisas
Nivel de explicación	Global	PDP, ICE, Feature Importance	Comprensión general del comportamiento del modelo
	Local	LIME, SHAP (valores	Explicación de decisiones

	individuales)	específicas

La elección de una técnica de XAI depende del tipo de modelo, la naturaleza de los datos, el objetivo específico de la explicación y las necesidades del usuario final [20]. En el contexto de la detección de somnolencia, donde se emplean CNNs para análisis de imágenes faciales, las técnicas basadas en gradiente como Grad-CAM y Score-CAM son particularmente apropiadas por su capacidad de generar mapas de calor visuales que indican las regiones faciales determinantes en la predicción.

2.1.5 Grad-CAM (Gradient-weighted Class Activation Mapping)

Es una técnica de explicabilidad visual post-hoc muy usada para modelos basados en Redes Neuronales Convolucionales (CNN) especialmente en tareas de visión por computadora, en donde su objetivo es hacer más transparentes las decisiones de estos modelos. Es una versión generalizada de CAM, que supera sus limitaciones al no requerir modificar la arquitectura del modelo. Grad-CAM se puede aplicar a una amplia variedad de modelos, incluso con capas totalmente conectadas, sin necesidad de cambios ni reentrenamiento, y se aplica en tareas como clasificación de imágenes, subtitulado y respuesta visual a preguntas [5].

Grad-CAM crea "explicaciones visuales" en forma de mapas de localización gruesos, o *heatmaps*. Estos resaltan las regiones importantes en la imagen de entrada que contribuyeron a la predicción de un concepto objetivo. La técnica usa los gradientes que fluyen hacia la última capa convolucional del modelo. Estos gradientes ayudan a calcular que tan relevantes son ciertas zonas de la imagen. Luego, se combinan linealmente con los mapas de activación para producir el mapa de calor final, al que se aplica una función ReLU para conservar solo las contribuciones positivas. A diferencia de CAM, que solo funcionaba en ciertas arquitecturas y requería modificar la red, Grad-CAM se puede aplicar a una amplia variedad de modelos, incluso con capas totalmente conectadas, sin necesidad de cambios ni reentrenamiento [5].

Una de sus principales ventajas es su bajo costo computacional en comparación con SHAP [22]. Sin embargo, puede ser sensible a pequeñas variaciones en los gradientes y no ofrece el mismo nivel de detalle en la contribución de características individuales.

Además, su interpretabilidad es principalmente cualitativa, ya que indica qué regiones son relevantes, lo que requiere una evaluación subjetiva por parte de expertos [22], [23].

2.1.6 Score-CAM

A diferencia de los métodos basados en gradientes, Score-CAM es un método novedoso de explicabilidad visual post-hoc para CNN que se distingue por no depender de los gradientes, a diferencia de Grad-CAM [8], [23].

La determinación del peso de cada mapa de activación se basa en su puntuación de pase hacia adelante (forward passing score) en la clase objetivo. La idea central es el concepto de "Aumento de Confianza" (CIC). El proceso comienza extrayendo los mapas de activación de una capa convolucional seleccionada. Luego, cada mapa se sobre muestrea, normaliza y se usa como máscara sobre la imagen original, que se reintroduce en la red para obtener una puntuación específica. Estas puntuaciones se normalizan, generalmente con softmax, para mejorar la discriminación entre clases. Finalmente, se combinan linealmente los mapas de activación con estos pesos y se aplica ReLU, generando así el mapa final que resalta las regiones más relevantes para la predicción [8].

Este enfoque presenta varias ventajas frente a métodos basados en gradientes como Grad-CAM y Grad-CAM++: no depende de gradientes, lo que evita ruido y saturación; asigna pesos más intuitivos basados en el “Aumento de Confianza”; produce mapas más suaves y discriminativos; ofrece mejor rendimiento visual y cuantitativo; y pasa la prueba de cordura, demostrando su fiabilidad. Además, es útil para depurar modelos. Sin embargo, su principal desventaja es el mayor costo computacional, ya que requiere múltiples pasos hacia adelante para calcular los pesos [12], [13].

2.1.7 LIME (Local Interpretable Model-agnostic Explanations)

Esta técnica es ampliamente adoptada en Inteligencia Artificial Explicable (XAI), reconocida por ofrecer explicaciones locales para modelos complejos. Su principal fortaleza radica en su capacidad para aplicarse a cualquier clasificador o regresor sin necesidad de conocer la estructura interna del modelo, operando después del entrenamiento y enfocándose en la relación entre entradas y salidas para generar explicaciones específicas de predicciones individuales [6], [24]**.**

El funcionamiento de LIME sigue una secuencia estructurada de cuatro pasos. Primero, durante la generación de características, la técnica adapta la instancia de entrada

según el tipo de dato: segmentando las imágenes en superpíxeles, tokenizando texto o codificando variables en datos tabulares. En la fase de generación de muestras, crea un conjunto de instancias perturbadas alrededor de la instancia original y obtiene sus predicciones mediante el modelo. Durante la atribución de características, entrena un modelo interpretable (usualmente lineal) para aproximar el comportamiento del modelo original localmente, ponderando las muestras según su cercanía a la instancia de interés, donde los coeficientes indican la influencia de cada característica. Finalmente, presentan los resultados de forma comprensible palabras clave resaltadas en texto, atributos importantes en tablas o superpíxeles marcados visualmente en imágenes [24].

A pesar de sus beneficios, LIME presenta desafíos importantes como inestabilidad en sus explicaciones y alta demanda computacional debido a la generación de múltiples muestras perturbadas [24]. Aún así, ha demostrado su versatilidad en aplicaciones especializadas como EmoLIME para reconocimiento de emocione en señales de voz, CoughLIME para detección de COVID-19 mediante el análisis de tos y en áreas tradicionales como visión por computadora o procesamiento de lenguaje natural [25].

2.1.8 SHAP (SHapley Additive exPlanations)

Complementando el enfoque de perturbación de LIME, SHAP es una técnica basada en los valores de Shapley de la teoría de juegos cooperativos, cuyo objetivo es asignar una puntuación de importancia a cada característica de entrada, cuantificando su contribución a la predicción del modelo [7], [26]. Los valores de Shapley dividen equitativamente una recompensa (en este caso, la predicción) entre características (jugadores), considerando todas las combinaciones posibles de características. De esta manera el valor SHAP de una característica refleja su cambio promedio en la predicción al incluirla en todas las coaliciones posibles.

Una de las fortalezas de SHAP radica en que cumple propiedades matemáticas como consistencia y precisión local, donde la suma de los valores SHAP iguala exactamente la diferencia entre la predicción del modelo para una instancia y la predicción de referencia, permitiendo una descomposición clara e interpretable de las predicciones [26]. Como método agnóstico del modelo, SHAP proporciona explicaciones locales sobre como cada característica de una instancia de prueba influye en la decisión del modelo [27].

Sin embargo, calcular estos valores tiene un alto costo computacional debido al número exponencial de evaluaciones requeridas, por lo que se utilizan estimadores como KernelSHAP, implementado en la librería SHAP, mientras que para modelos basados en árboles existen métodos más eficientes [27]. Además, presenta desafíos para usuarios no técnicos, ya que sus explicaciones basadas en atribución de características son sensibles a la interpretación, donde diferentes técnicas de ingeniería pueden alterar los valores de importancia y ocultar discriminación [26], [28]. En aplicaciones como HAR, SHAP es menos efectivo que métodos como Grad-CAM para capturar dinámicas espaciales y temporales [22].

2.1.9 Métricas específicas para evaluación de explicabilidad

Mientras que la sección 2.1.1 presentó los principios conceptuales de XAI (transparencia, interpretabilidad y explicabilidad), a continuación, se detallan las métricas específicas para evaluar cuantitativamente estos principios en la práctica.

La evaluación de la Inteligencia Artificial Explicable (XAI) representa un gran desafío, principalmente por la falta de métricas estandarizadas y confiables. Esta carencia no solo reduce el valor práctico y la fiabilidad de las explicaciones, sino que también dificulta el cumplimiento normativas y la comparación entre métodos al no existir una "verdad fundamental" [29].

Para abordar este problema, las métricas de evaluación de XAI se organizan considerando distintos criterios, entre los que destacan [30]:

- **Fidelidad (Faithfulness):** Mide qué tan bien una explicación representa el proceso de decisión real del modelo, lo cual se puede comprobar aleatorizando sus parámetros y comparando con modelos de caja blanca, o midiendo cómo cambia la salida al eliminar o perturbar características importantes [31].

- **Robustez (Robustness):** Evalúa la estabilidad y consistencia de las explicaciones bajo diferentes entradas, incluyendo ataques adversarios y perturbaciones [31].

- **Localización (Localization):** Evalúa la capacidad de la explicación para señalar con precisión que regiones o características relevantes en los datos de entrada fueron los más influyentes en la decisión del modelo [30].

Dado que ningún método XAI puede ser perfecto en todas las métricas al mismo

tiempo, para una correcta evaluación se requiere analizar cada propiedad por separada y encontrar un equilibrio adecuado entre todas ellas [31].

Habiendo establecido los fundamentos de XAI y sus técnicas principales, es necesario comprender las arquitecturas de Deep Learning sobre las cuales se aplicarán estos métodos de explicabilidad.

2.2 Deep Learning y Redes Neuronales Convolucionales

La evolución reciente en inteligencia artificial ha dado lugar a un cambio paradigmático en cómo se abordan los problemas de aprendizaje automático. Este cambio, de particular relevancia para la detección de somnolencia, redefine las prioridades en el desarrollo de sistemas inteligentes.

2.2.1 Fundamentos de las redes neuronales convolucionales (CNN)

Las Redes Neuronales Convolucionales (CNNs) son una arquitectura de Deep Learning inspirada en el córtex visual animal, diseñadas específicamente para el procesamiento de imágenes [32]. Su diseño se fundamenta en tres principios clave que las distinguen de las redes neuronales tradicionales:

- **Conectividad local**: Las neuronas se conectan y analizan únicamente pequeñas regiones de la capa anterior, imitando el funcionamiento de los campos receptivos visuales en sistemas biológicos.
- **Pesos compartidos**: Los filtros aprendidos se replican en todo el campo visual para detectar características sin importar su posición en la imagen, lo que permite invariancia a la traslación.
- **Pooling**: Reduce progresivamente las dimensiones espaciales, disminuyendo la carga computacional y proporcionando invariancia a pequeñas traslaciones [32].

La estructura de una CNN transforma el volumen de entrada (imagen) en un volumen de salida (puntuaciones de clase) a través de una secuencia de capas especializadas. El bloque fundamental es la capa convolucional, que desliza filtros sobre la imagen para crear mapas de características, usualmente activados con la función ReLU. A continuación, las capas de pooling comprimen estos mapas reduciendo su dimensionalidad espacial, y finalmente, las capas completamente conectadas reciben un vector aplanado de las características de alto nivel para realizar la clasificación final [33].

Ventajas de las CNNs para procesamiento visual

Esta arquitectura jerárquica hace que las CNNs sean superiores a las redes neuronales estándar para tareas de visión por computadora por múltiples razones. Primero, requieren significativamente menos parámetros que las redes completamente conectadas debido al uso de pesos compartidos, lo que hace su entrenamiento más eficaz y reduce el riesgo de sobreajuste. Segundo, aprovechan inherentemente la estructura espacial de las imágenes, asumiendo que los píxeles cercanos están correlacionados, lo cual es una característica fundamental de los datos visuales. Esta capacidad para extraer automáticamente características visuales jerárquicas las hace ideales para aplicaciones como la detección de somnolencia, donde pueden identificar progresivamente patrones como el estado de los ojos, la apertura de la boca y la postura de la cabeza [34].

2.2.2 Métricas de rendimiento: precisión, recall, F1-score

Para evaluar los modelos de clasificación en tareas como la detección de somnolencia, donde los datos suelen estar desequilibrados, es fundamental usar métricas que brinden una visión más completa que la exactitud global. Su cálculo se realiza a partir de cuatro resultados de una clasificación binaria [35]: los Verdaderos Positivos (TP) y los Verdaderos Negativos (TN), que representan las predicciones correctas; y los errores, igualmente importantes, conformados por los Falsos Positivos (FP) o errores Tipo I, y los Falsos Negativos (FN) o errores Tipo II.

A partir de estos cuatro valores, se calculan las métricas de rendimiento:

Tabla 2. Métricas de Rendimiento para Modelos de Clasificación [35]

Métrica	Fórmula	Descripción
Precisión (Precision)	$P = \frac{TP}{TP + FP}$	Mide la exactitud de las predicciones positivas, siendo clave cuando el costo de un falso positivo es alto.

Recall (Exhaustividad / Sensibilidad)	$R = \frac{TP}{TP + FN}$	Mide la capacidad para identificar todos los casos positivos. Es vital cuando el costo de un falso negativo es alto.
F1-score (F1-score)	$F1 = 2 \times \frac{P \times R}{P + R}$	Media armónica de la precisión y el recall. Ideal para conjuntos de datos desequilibrados.

A menudo, existe una relación inversa entre precisión y recall [35]. En detección de somnolencia, un falso negativo es más peligroso que un falso positivo, por lo tanto, se priorizar un alto recall.

2.2.3 Aplicaciones de Deep Learning en la detección de somnolencia

El Deep Learning ha revolucionado la detección de somnolencia en conductores, mediante la visión por computadora, mejorando la seguridad vial [31]. La capacidad de las redes neuronales profundas al analizar características visuales, especialmente el estado de los ojos, gracias a la capacidad de las redes neuronales profundas para aprender patrones complejos directamente de imágenes y videos.

Las principales técnicas implementadas incluyen:

- **Detección de Rostros y Puntos Característicos:** CNNs como MTCNN detectan rostros y localizan puntos clave alrededor de los ojos(landmarks) [31].
- **Relación de Aspecto del Ojo (EAR - Eye Aspect Ratio):** Esta métrica cuantifica el grado de apertura o cierre del ojo a partir de los puntos de referencia oculares. Un valor bajo indica ojos cerrados y un valor alto indica abiertos [23], [31].
- **PERCLOS (Percentage of Eyelid Closure Over the Pupil Over Time):** Siguiendo el estándar psicofisiológico detallado en la sección 2.5.1, se mide el porcentaje de cierre de párpados en un período determinado. Un valor >0.15 en 150 fotogramas se utiliza como indicador de fatiga [36].
- **Frecuencia de Parpadeo y Duración del Cierre Ocular:** Un aumento en la frecuencia y cierre de parpadeo indican somnolencia [37].

La implementación de estos sistemas en tiempo real requiere una alta eficiencia computacional y grandes conjuntos de datos etiquetados, ampliados mediante aumento de datos para mejorar la robustez del modelo [31].

2.3 Paradigmas Contemporáneos en el Desarrollo de Sistemas Inteligentes

La evolución de la inteligencia artificial aplicada a problemas de visión por computadora, como la detección de somnolencia, ha transitado por diversas etapas filosóficas y técnicas. Para comprender la justificación detrás de las estrategias de preprocesamiento avanzado propuestas en esta investigación, resulta necesario analizar el cambio de paradigma que se está produciendo en la comunidad científica: el paso de una visión centrada en el modelo a una centrada en los datos.

2.3.1 La Transición del Model-Centric al Data-Centric AI

Históricamente, la investigación en Machine Learning y Deep Learning ha estado dominada por el paradigma conocido como *Model-Centric AI*. Este enfoque, el conjunto de datos se considera una entidad estática, fija e inmutable. Los ingenieros e investigadores dedican la inmensa mayoría de sus recursos y tiempo a iterar sobre la arquitectura del modelo, ajustando hiperparámetros, diseñando nuevas funciones de pérdida, profundizando las capas de las redes neuronales o modificando los algoritmos de optimización para exprimir mejoras marginales en el rendimiento predictivo [38], [39].

Si bien esta estrategia ha sido responsable de avances monumentales en la última década, autores influyentes como Andrew Ng han argumentado recientemente que este enfoque ha alcanzado un punto de rendimientos decrecientes en muchas aplicaciones prácticas. En escenarios del mundo real, donde los datos suelen ser ruidosos, escasos o estar sujetos a sesgos de dominio, la simple sofisticación del modelo matemático a menudo conduce al sobreajuste (overfitting) en lugar de a una generalización robusta [40].

En respuesta a estas limitaciones, surge el paradigma de Data-Centric AI (Inteligencia Artificial Centrada en Datos). Este enfoque propone mantener la arquitectura del modelo relativamente fija (o utilizar arquitecturas estándar probadas) y centrar el esfuerzo iterativo en la mejora sistemática de la calidad de los datos [41]. La premisa fundamental es que los datos no son simplemente un insumo pasivo, sino un "código" que debe ser diseñado, depurado y optimizado.

En el contexto de la detección de somnolencia, adoptar un enfoque *Data-Centric* implica reconocer que la calidad de la información visual que se presenta a la red neuronal es más determinante que la profundidad de la red misma. Si las imágenes de entrada contienen ruido contextual irrelevante (como el fondo del vehículo, patrones en la ropa del conductor o condiciones de iluminación variables), el modelo puede verse obligado a aprender patrones incorrectos.

2.3.2 El Fenómeno "Clever Hans" y el Aprendizaje de Atajos

Uno de los argumentos más potentes a favor de un preprocesamiento riguroso y guiado por explicabilidad es la existencia del fenómeno conocido como "Clever Hans" en los modelos de Deep Learning. Este término proviene de la historia de un caballo a principios del siglo XX que supuestamente podía realizar operaciones aritméticas, pero que en realidad había aprendido a leer las señales corporales involuntarias de su entrenador para dar la respuesta correcta, sin entender en absoluto las matemáticas [42].

En el ámbito de la visión por computadora, el efecto *Clever Hans* se refiere a la tendencia de las redes neuronales profundas a aprender correlaciones espurias o atajos (*shortcuts*) presentes en los datos de entrenamiento, en lugar de aprender las características causales y semánticas del objeto de interés. Lapuschkin et al. [43], en un estudio seminal, demostraron mediante técnicas de XAI que modelos de clasificación de imágenes de estado del arte, que reportaban una precisión excepcional, basaban sus decisiones en artefactos irrelevantes. Por ejemplo, clasificaban imágenes como "caballo" detectando una marca de agua de copyright específica que aparecía frecuentemente en las fotos de caballos del dataset Pascal VOC, ignorando por completo la morfología del animal.

En el contexto de la detección de somnolencia, el efecto Clever Hans es devastador. Un modelo entrenado con datos de simuladores podría aprender a asociar el "fondo gris" del simulador con la clase "somnoliento" y el "fondo de calle real" con la clase "alerta". Si se despliega este modelo en un coche real, fallará catastróficamente porque el fondo ha cambiado, o peor aún, detectará somnolencia solo cuando el coche pase por un túnel (fondo oscuro) [44].

2.4 Ingeniería de Datos Guiada por XAI (XAI-Guided Data Engineering)

La convergencia entre la necesidad de datos de alta calidad (*Data-Centric AI*) y la necesidad de evitar correlaciones espurias (*Clever Hans mitigation*) ha dado lugar a una

nueva metodología: la Ingeniería de Datos Guiada por XAI (*XAI-Guided Data Engineering*). Tradicionalmente, la Inteligencia Artificial Explicable (XAI) se ha utilizado como una herramienta de auditoría *post-hoc*, aplicada al final del ciclo de desarrollo para "explicar" lo que el modelo ha aprendido. Sin embargo, este nuevo enfoque propone utilizar las explicaciones durante la fase de desarrollo para guiar activamente el diseño y refinamiento de los datos [45].

El ciclo de trabajo en este paradigma es iterativo y se compone de estas fases:

1. **Diagnóstico Visual:** Se entrena un modelo base y se utilizan técnicas de atribución de características (como Grad-CAM, LIME o SHAP) para generar mapas de saliencia. Estos mapas revelan las regiones de la imagen que más influyen en la predicción [45].

2. **Identificación de Sesgos:** Se analizan los mapas de saliencia para detectar patrones de atención incorrectos. Por ejemplo, si los mapas de calor destacan sistemáticamente el fondo de la imagen o accesorios irrelevantes, se identifica un sesgo de atención [46].

3. **Intervención en el Preprocesamiento:** Basándose en el diagnóstico, se diseñan técnicas de preprocesamiento específicas para suprimir la información irrelevante o resaltar la relevante. Esto puede incluir técnicas como el recorte adaptativo (*adaptive cropping*), el desenfoque selectivo (*selective blurring*) o el enmascaramiento de regiones (*masking*) [46].

4. **Reentrenamiento y Validación:** Se entrena un nuevo modelo con los datos modificados y se verifica nuevamente con XAI que la atención se haya desplazado hacia las características causales correctas (los ojos y la boca en el caso de la somnolencia) [47].

Este enfoque permite transformar modelos de "caja negra" propensos a errores en modelos robustos y alineados con el conocimiento humano, utilizando la explicabilidad no solo para entender, sino para mejorar el sistema [44], [47].

2.5 Fundamentos Fisiológicos y Computacionales de la Atención Visual

2.5.1 El Estándar PERCLOS y la Relevancia Biométrica

En el dominio de la seguridad vial y la psicofisiología, el indicador PERCLOS (Percentage of Eye Closure) se ha establecido como el "estándar de oro" (*gold standard*) para la detección objetiva de la somnolencia. Numerosos estudios, incluyendo la revisión exhaustiva de Abe et al., confirman que los cambios en la dinámica oculomotora específicamente la duración y frecuencia del cierre palpebral, así como la velocidad de parpadeo son los predictores más fiables del deterioro de la vigilancia. Indicadores secundarios como la frecuencia de bostezos y las expresiones faciales de fatiga también aportan información valiosa, pero la región ocular sigue siendo la fuente primaria de señal [36].

Esto implica que, desde una perspectiva de teoría de la información, la "señal" relevante para la clasificación de somnolencia está espacialmente concentrada en la región periocular y bucal, mientras que el resto de la imagen facial y el fondo constituyen principalmente "ruido".

2.5.2 Simulación de la Visión Foveal y Supresión de Fondo

El sistema visual humano opera mediante un mecanismo de visión foveal: la retina captura con alta resolución solo una pequeña área central (la fóvea) hacia donde se dirige la mirada, mientras que la visión periférica tiene una resolución mucho menor y sirve principalmente para detectar movimiento y contexto global. Este mecanismo permite al cerebro procesar escenas complejas de manera eficiente, filtrando irrelevancias [48].

En Deep Learning, se ha demostrado que simular este comportamiento de procesar con alta fidelidad las regiones de interés y degradando la periferia, puede aumentar significativamente la robustez de los modelos ante ataques adversarios y perturbaciones naturales. Técnicas como el desenfoque Gaussiano (*Gaussian Blur*) actúan como filtros de paso bajo que eliminan las frecuencias espaciales altas (bordes nítidos, texturas finas) de las regiones no esenciales. Al aplicar este desenfoque al fondo de las imágenes de entrenamiento, se eliminan las texturas que podrían inducir correlaciones espurias (como el tejido del asiento del coche), forzando a la red neuronal a concentrar su capacidad de aprendizaje en las características biométricas que permanecen nítidas. Este proceso se conoce como supresión de fondo (*background*

suppression) y es una técnica validada para mejorar la generalización en tareas de clasificación y detección de objetos [48], [49]. En la presente investigación, este principio fundamenta la propuesta del 'Escenario B', donde se implementa un enmascaramiento dinámico y desenfoque periférico para forzar al modelo a extraer características exclusivamente de las regiones biométricas clave (ojos y boca).

2.6 Detección de Somnolencia en Conductores

2.6.1 Definición y características fisiológicas de la somnolencia

La somnolencia es un estado complejo y multifacético que se distingue por una alteración del mecanismo normal de excitación, representando una transición entre la vigilia y el sueño.67 En el contexto de la conducción, se define como la acción de conducir cuando se presentan síntomas de fatiga, somnolencia o incapacidad para mantener el estado de alerta, pudiendo ser causada por factores como la falta de sueño o largas horas de conducción [50]. Este estado se manifiesta a través de diversas características fisiológicas que reflejan cambios en la actividad cerebral y corporal, observándose claramente en el comportamiento ocular donde las ondas cerebrales cambian de patrones de vigilia (ondas beta) a patrones asociados con el sueño ligero (ondas alfa y theta), traduciéndose en indicadores externos como la alteración del cierre de los párpados y la reducción del tiempo de respuesta [51].

2.6.2 Impacto de la somnolencia en la seguridad vial

La somnolencia al volante es un problema crítico de seguridad vial global, responsable de un número alarmante de muertes, lesiones y pérdidas económicas cada año [34], [35], [36]. Contribuye aproximadamente 1.19 millones de fallecimientos anuales por accidente de tráfico reportados en 2021 [11], como lo demuestran las estadísticas estadounidenses donde entre 2017 y 2021, el 17,6% de todos los accidentes fatales involucraron conductores somnolientes, resultando en aproximadamente 29,834 muertes [36], [38].

El análisis demográfico revela que el porcentaje accidentes es alto en jóvenes, el mayor número se concentra en el grupo de 21-34 años, con una notable predominancia de hombres. Aproximadamente un tercio había consumido alcohol, y los siniestros ocurren principalmente entre las 11:00 PM y 2:59 AM con más frecuencia en carreteras rurales secundarias. Esto se debe a que la somnolencia compromete la atención, el tiempo

de reacción y la capacidad de decisión, lo que multiplica de cuatro a seis veces más el riesgo de accidentarse que un conductor alerta [54].

Ante esta problemática, el desarrollo de sistemas de asistencia al conductor avanzados (ADAS) es se ha vuelto una prioridad. Estos sistemas utilizan cámaras y sensores para monitorear indicadores visuales y fisiológicos, alertando al conductor antes de eventos críticos para reducir las lesiones y fatalidades [30].

2.7 Metodología CRISP-DM Aplicada a XAI

CRISP-DM (Cross-Industry Standard Process for Data Mining) es un modelo de proceso estándar ampliamente adoptado para proyectos de minería de datos y aprendizaje automático, que proporciona un enfoque estructurado y no propietario para planificar y ejecutar proyectos de ciencia de datos [10]. Su diseño iterativo y flexible lo hace particularmente adecuado para proyectos que integran Inteligencia Artificial Explicable (XAI), donde la comprensión del modelo es tan importante como su rendimiento predictivo.

La metodología se estructura en seis fases interrelacionadas que permiten abordar sistemáticamente el desarrollo de sistemas de IA: Comprensión del Negocio, Comprensión de Datos, Preparación de Datos, Modelado, Evaluación y Despliegue [10]. En el contexto de XAI, cada fase puede incorporar consideraciones específicas de explicabilidad y transparencia [55].

Tabla 3. Fases de CRISP-DM y sus consideraciones en proyectos de XAI [10], [55]

Fase	Enfoque Principal	Consideraciones de XAI
1. Comprensión del Negocio	Definir objetivos y requisitos del proyecto	Establecer los objetivos de rendimiento y explicabilidad, considerando transparencia, ética y regulaciones.
2. Comprensión de Datos	Explorar, describir y verificar calidad de datos	Identificar sesgos potenciales y evaluar representatividad

3. Preparación de Datos	Seleccionar, limpiar y construir el dataset final	Aplicar transformaciones que preserven interpretabilidad
4. Modelado	Seleccionar técnicas y construir modelos	Equilibrar complejidad del modelo con explicabilidad requerida
5. Evaluación	Verificar que el modelo cumple objetivos	Evaluar tanto rendimiento predictivo como calidad de explicaciones
6. Despliegue	Planificar implementación y mantenimiento	Documentar decisiones del modelo para auditoría y monitoreo continuo

La naturaleza iterativa de CRISP-DM permite que los hallazgos en fases avanzadas informen mejoras en fases anteriores, facilitando un ciclo de refinamiento continuo [10]. Esta característica es particularmente valiosa en proyectos de XAI, donde los análisis de explicabilidad pueden revelar la necesidad de modificaciones en el preprocesamiento de datos o la arquitectura del modelo.

2.8 Trabajos relacionados

En la reciente investigación [56] desarrollaron sistemas de detección de emociones y somnolencia del conductor utilizando CNN y técnicas de Inteligencia Artificial Explicable (XAI). El sistema de detección de somnolencia se basó en el entrenamiento de varios modelos, incluyendo LeNet, con el conjunto de datos UTA-RLDD, compuesto por videos procesados en fotogramas faciales, y dividido por sujeto para garantizar una validación robusta y generalización a sujetos no vistos. Aunque AlexNet alcanzó la mayor exactitud (71.84%), se optó por LeNet como modelo final debido a su mejor equilibrio entre exactitud (67.97%) y tasa de falsos positivos (32%). La aplicación de XAI a través de XRAI (Extended Randomized Input Sampling for Explanation) evidenció que, mientras el modelo se enfoca correctamente en los ojos para clasificaciones precisas, también considera características irrelevantes (bordes de gafas) y puede generar clasificaciones erróneas por etiquetado inadecuado del dataset. Este hallazgo proporciona evidencia empírica directa del fenómeno Clever Hans en sistemas de detección de somnolencia, donde el modelo aprende correlaciones espurias de

artefactos estáticos en lugar de biomarcadores causales de fatiga. Las limitaciones principales incluyen el análisis limitado a imágenes estáticas que no captura la naturaleza progresiva de la somnolencia, comprometiendo su efectividad en aplicaciones de tiempo real, y la ausencia de una metodología de reentrenamiento que corrija los sesgos identificados mediante XAI, dejando el ciclo de mejora incompleto.

La investigación [57] propone un modelo de detección de somnolencia en conductores basado en imágenes faciales mediante CNN y el conjunto de datos RLDD, que contiene aproximadamente 30 horas de videos RGB de 60 conductores. El preprocesamiento incluyó captura y orientación de fotogramas, detección facial con OpenCV, redimensionamiento a 224x224 píxeles, conversión a escala de grises y normalización. Se evaluaron tres arquitecturas CNN (VGG-16, DeXpression y una CNN genérica de tamaño mediano), donde DeXpression alcanzó el mejor rendimiento con más del 78% de exactitud en imágenes de escala de grises. Su principal aporte es la interpretación de las decisiones del modelo a través de múltiples técnicas XAI complementarias: Grad-CAM, Integrated Gradients, Layer-wise Relevance Propagation (LRP), LIME y Occlusion Sensitivity, que generaron explicaciones visuales permitiendo una validación cruzada de la atención del modelo. Los resultados destacan que Grad-CAM fue la técnica más efectiva en la identificación de zonas relevantes para imágenes correctamente clasificadas, enfocándose consistentemente en regiones oculares y bucales. Sin embargo, la efectividad de todas las técnicas se redujo en imágenes mal clasificadas, fallando en enfocarse en áreas de mayor importancia. Las limitaciones incluyen la necesidad de mayor cantidad de datos de entrenamiento, preferiblemente de conductores reales en lugar de escenarios simulados, la ausencia de una estrategia de corrección basada en los hallazgos XAI para reentrenar el modelo eliminando los sesgos identificados, y la falta de validación de su aplicabilidad práctica en sistemas de tiempo real con restricciones computacionales.

La propuesta [58] introduce un método de detección de somnolencia mediante análisis de la región ocular como indicador temprano de fatiga, aprovechando MediaPipe para la extracción de la región de interés (ROI) correspondiente a los ojos. Se evaluaron tres redes neuronales profundas (Inception V3, VGG16 y ResNet50V2) con videos del dataset NITYMED, que contiene grabaciones de conductores en entornos reales no controlados con diversas condiciones de iluminación. El preprocesamiento incluyó extracción de fotogramas, detección de 468 puntos faciales con MediaPipe y una técnica

mejorada para corrección y selección de ROI ocular que minimiza la pérdida de información por movimiento de cabeza. Este proceso generó un dataset de 6800 imágenes distribuidas en 75% entrenamiento, 15% validación y 15% pruebas, redimensionadas a 112×112 píxeles, normalizadas de 0-255 a 0-1, con técnicas de aumento aplicadas para prevenir sobreajuste. La arquitectura CNN para clasificación binaria adaptó las salidas de redes pre-entrenadas mediante capas densas y dropout, configurándose para activarse cuando la probabilidad "Drowsy" superara 95% y los ojos permanecieran cerrados más de 300 milisegundos. ResNet50V2 demostró ser la arquitectura más efectiva con 99.71% de precisión, validado mediante Grad-CAM que reveló su concentración en la región ocular, confirmando que el modelo aprende de características relevantes y no de artefactos del fondo. Este trabajo ejemplifica el enfoque data-centric donde el preprocesamiento inteligente (extracción de ROI mediante MediaPipe) combinado con arquitecturas estándar logra resultados superiores a arquitecturas complejas sin preprocesamiento especializado. La principal limitación es su dependencia de la detección precisa de puntos faciales, ya que su rendimiento se vería afectado cuando el rostro no sea completamente visible debido varios factores.

Tang et al. [59] propusieron una red convolucional multiescala guiada por atención (MSCNN-CAM) para abordar la incapacidad de las CNN tradicionales de procesar simultáneamente características de grano fino (micro-clausura palpebral) y globales (postura de cabeza). Su arquitectura integra ramas convolucionales paralelas con kernels de diferentes tamaños y un mecanismo de atención de canal (CAM) que recalibra adaptativamente el peso de cada canal de características, asignando valores cercanos a cero a información de fondo y valores altos a bordes dinámicos en ojos y boca. Evaluado en el dataset SEED-VIG, alcanzaron una exactitud del 95.6%, superando a ResNet-50 y VGG-16 por 3-5%. Como técnica XAI, utilizaron mapas de calor que demostraron que, tras el entrenamiento guiado por atención, la activación se concentraba exclusivamente en regiones perioculares y bucales, ignorando variaciones de iluminación. Este trabajo valida que la focalización puede ser aprendida por la red mediante mecanismos de atención integrados, eliminando la necesidad de recorte manual extensivo. Sin embargo, su limitación principal radica en la alta complejidad computacional de las ramas multiescala, lo que podría dificultar su implementación en sistemas embebidos o de recursos limitados, sugiriendo la necesidad de estrategias híbridas más eficientes.

Sellal et al. [60] formalizaron la metodología de "Ingeniería de Características Guiada por XAI" implementando un ciclo iterativo de refinamiento para reconocimiento de estilo de conducción. Su enfoque consistió en entrenar un modelo base con todas las características disponibles, realizar auditoría SHAP global para identificar el impacto de cada característica, detectar y podar características espurias (identificaron que el modelo dependía de "ruido de sensores" como vibraciones irrelevantes que correlacionaban espuriamente), y finalmente reentrenar el modelo con el conjunto limpio. Utilizando SHAP (Shapley Additive Explanations) como técnica XAI central, lograron que el modelo "podado" alcanzara 92% de exactitud con Random Forest, superando al modelo original que utilizaba todas las características, demostrando que la eliminación selectiva de información puede mejorar la generalización al forzar el aprendizaje de patrones causales reales. El trabajo valida que la poda de características visuales basada en evidencia XAI de correlaciones espurias constituye una estrategia científicamente fundamentada para mitigación de sesgos. La limitación principal es que el enfoque requiere múltiples ciclos de entrenamiento completo, lo que incrementa significativamente el costo computacional y el tiempo de desarrollo, y depende críticamente de la calidad e interpretabilidad de las explicaciones SHAP que pueden ser inestables en datasets pequeños.

Mersha et al. [61] propusieron un pipeline innovador donde las explicaciones del modelo guían las técnicas de aumento de datos de forma selectiva y contextual. Su metodología consiste en usar XAI para identificar regiones "importantes" (ojos) versus "no críticas" (fondo), y luego aplicar técnicas de aumento (ruido, rotación, desenfoque) de manera selectiva: aplicando ruido intenso a zonas no importantes para forzar al modelo a no depender de ellas preservando características críticas, o viceversa, perturbando características críticas para generar ejemplos adversarios que robustezcan la frontera de decisión. Utilizando mapas de saliencia como técnica XAI guía, demostraron mejoras significativas en precisión (hasta 8.1% en tareas de NLP, extrapolable a visión) y reducciones drásticas en el sobreajuste a características espurias. La propuesta establece una metodología avanzada de reentrenamiento que combina recorte de imágenes con aplicación de Aumento de Datos Guiado por XAI, específicamente ruido gaussiano o "CutOut" en zonas previamente identificadas como distractores (marcos de gafas, reposacabezas) durante el proceso de entrenamiento. Las limitaciones principales incluyen: la necesidad de análisis XAI preliminar que añade complejidad computacional al pipeline de entrenamiento, la dependencia de la calidad de los mapas de saliencia que

pueden ser ruidosos o inconsistentes, y la falta de validación extensiva en tareas de visión por computadora en tiempo real donde el aumento agresivo podría introducir artefactos que degraden la calidad de señales temporales críticas como la dinámica del parpadeo.

La propuesta [62] desarrolla el Multi-Attention Fusion Drowsy Driving Detection Model (MAF), un modelo que integra CNN con mecanismos de atención múltiples para mejorar la clasificación de somnolencia en conductores bajo condiciones de oclusión facial parcial y baja iluminación. El MAF emplea una red troncal para extracción de características y módulos de atención espacial (LANets) paralelos que, combinados con desactivación aleatoria del mapa de atención durante el entrenamiento, enfocan el modelo en regiones faciales cruciales ignorando artefactos oclusivos, mientras que un Módulo de Mejora de Características Locales Aprendibles refina estas características mediante mecanismos de atención cruzada. Se evaluó en un dataset propio que incluye conductores con gafas y videos diurnos/nocturnos capturando escenarios adversos realistas, además del dataset RLDD de referencia, logrando 96.8% de exactitud en el conjunto propio y 99.11% en RLDD. La visualización mediante Grad-CAM como técnica XAI demostró que el modelo se enfoca consistentemente en características prominentes como ojos y boca incluso bajo oclusiones como mascarillas o gafas de sol, confirmando que los mecanismos de atención múltiple permiten al modelo discriminar entre artefactos oclusivos (que deben ignorarse) y biomarcadores de fatiga (que deben priorizarse). Este trabajo aborda directamente el problema de artefactos estáticos (gafas, accesorios) mediante soluciones arquitecturales, demostrando efectividad en escenarios donde la oclusión es inherente y no puede eliminarse mediante preprocesamiento. Sin embargo, las limitaciones incluyen la falta de detalles sobre inferencia en tiempo real o implementación en hardware embebido, la ausencia de análisis de complejidad computacional comparativa con arquitecturas más simples, y no se evalúa si un enfoque data-centric (recorte adaptativo de ROI) podría lograr resultados similares con menor costo computacional que los múltiples módulos de atención paralelos.

CAPÍTULO III

3. MATERIALES Y MÉTODOS

Enfoque Metodológico y Estándar de Proceso

Para estructurar el ciclo de vida del desarrollo, se ha seguido la metodología CRISP-DM. Este estudio se enfoca específicamente en la ejecución de las fases 2 a 5 (desde la Comprensión de Datos hasta la Evaluación del Modelo), integrando la explicabilidad como un requisito transversal que informa cada etapa del proceso. La Fase 1 fue abordada en el Capítulo I y la Fase 6 queda fuera del alcance experimental de esta investigación.

Tabla 4. Fases de CRISP-DM implementadas en este estudio

Fase CRISP-DM	**Actividad Experimental Específica**	**Herramientas y Tecnologías**
Fase 2: Comprensión de Datos	Análisis exploratorio del Drowsiness Prediction Dataset, verificación de balance de clases y evaluación de sesgos	Pandas, Matplotlib, OpenCV
Fase 3: Preparación de Datos	Implementación de dos pipelines contrastantes: Escenario A (Inyección de Malla MediaPipe) vs. Escenario B (Focalización Biométrica)	MediaPipe Face Mesh, NumPy, cv2
Fase 4: Modelado	Construcción y entrenamiento de una CNN secuencial con regularización mediante Dropout	TensorFlow 2.x, Keras, Adam Optimizer
Fase 5: Evaluación	Validación de rendimiento (Accuracy, Recall, F1-Score) y auditoría de explicabilidad (Grad-CAM, Score-CAM, LIME, SHAP)	tf-explain, scikit-learn, LIME, SHAP

3.1 Entorno de Desarrollo e Infraestructura Computacional

Los modelos de redes neuronales convolucionales, y sobre todo los algoritmos de explicabilidad como LIME (basado en perturbación) o SHAP (valores de Shapley), necesitan una capacidad importante de cómputo vectorial. Por esta razón, optó por una infraestructura que minimizara los tiempos de entrenamiento y permitiera probar distintos hiperparámetros sin demoras.

3.1.1 Arquitectura de Hardware y Aceleración

El entorno experimental se desarrolló sobre la plataforma Kaggle Kernels [63], aprovechando su infraestructura basada en contenedores Docker para garantizar la portabilidad del código. Se trabajó con una NVIDIA Tesla P100 (PCIe) que cuenta con 16 GB de memoria VRAM dedicada. Esta arquitectura Pascal resultó fundamental para el proyecto por su capacidad de cómputo en precisión mixta, lo que acelera las operaciones de convolución matricial propias de las CNN. En particular, durante la fase de XAI, permitió procesar lotes grandes de imágenes perturbadas (necesarias para LIME) sin problemas de intercambio de memoria, lo que redujo el tiempo de generación de explicaciones de horas a apenas minutos.

3.1.2 Arquitectura de Hardware y Aceleración

El ecosistema de software se configuró cuidadosamente para garantizar que las librerías de visión artificial y el framework de entrenamiento funcionaran bien juntas, resolviendo conflictos de dependencias:

- **Lenguaje y Entorno:** Se utilizó Python 3.10+ como lenguaje base, aprovechando las mejoras de rendimiento de sus versiones recientes para procesar flujos de datos de forma más eficiente.
- **Deep Learning (TensorFlow 2.15.0 [64]):** Se eligió esta versión específica de TensorFlow/Keras por su soporte optimizado para ejecución "Eager", algo indispensable para calcular gradientes en tiempo real como los que requieren técnicas como Grad-CAM y GradientExplainer.
- **Visión Artificial (MediaPipe 0.10.15 [65]):** Para extraer características faciales se implementó la versión 0.10.15 de MediaPipe. Esta versión incluye mejoras importantes en la estabilidad temporal (anti-jitter) de la malla facial, algo fundamental para los algoritmos de recorte adaptativo que se proponen en esta investigación.

- **Procesamiento de Imágenes (OpenCV 4.8.0.76):** Se empleó opencv-python-headless para operaciones básicas como lectura de archivos, conversión de espacios de color (BGR a RGB) y aplicación de filtros Gaussianos. La variante "headless" se eligió deliberadamente para entornos de nube, eliminando dependencias de librerías gráficas del sistema (como X11) que suelen causar conflictos en contenedores Docker o entornos virtualizados.
- **Bibliotecas de Explicabilidad (XAI):** Se utilizaron tres herramientas principales: tf-explain [66] para implementar Grad-CAM y Score-CAM optimizados e integrados con TensorFlow; LIME [6] para generar explicaciones locales mediante modelos sustitutos lineales; y SHAP [7] para estimar la contribución de características usando teoría de juegos cooperativos con el GradientExplainer adaptado para redes profundas.
- **Gestión de Dependencias:** Para que todo el sistema fuera compatible, se fijaron dependencias críticas: Protobuf 3.20.3 (para que la serialización entre TensorFlow y MediaPipe funcionara correctamente) y NumPy < 2.0 (para evitar conflictos de ABI binaria).

3.2 Fase 2: Comprensión y Adquisición de los Datos

El insumo principal para entrenar los modelos predictivos es el Drowsiness Prediction Dataset, un conjunto de datos público disponible en Kaggle, creado por el investigador Rakibul [67]. Si bien este repositorio funciona como un agregador de datos, las imágenes provienen de bases de datos académicas validadas como UTA-RLDD [68] y YawDD [69]. Esta procedencia garantiza que el entrenamiento se sustente en capturas de video en entornos controlados que simulan la cabina de un vehículo.

3.2.1 Estructura y Distribución de Clases

El dataset se organiza en una estructura de directorios que facilita la carga de datos mediante generadores de flujo. Contiene imágenes faciales recortadas de fotogramas de video que capturan a conductores en distintos estados de alerta. Un aspecto favorable es que el dataset está perfectamente balanceado, lo que reduce el sesgo de clase y permite usar la exactitud (accuracy) como métrica de referencia inicial válida, aunque no la única.

En cuanto a los atributos visuales de los sujetos y el entorno, el dataset presenta las siguientes características técnicas:

- **Condiciones de iluminación:** Las imágenes corresponden predominantemente a iluminación diurna o luz artificial constante. Esto implica que el modelo está optimizado para trabajar en el espectro visible (RGB). La ausencia de imágenes con iluminación infrarroja (NIR) o condiciones de oscuridad total se reconoce como una delimitación del alcance del sistema actual.
- **Uso de accesorios:** El conjunto de datos incluye variabilidad en el uso de gafas de corrección (lentes transparentes), lo que permite al modelo aprender a ignorar los reflejos o monturas comunes. Sin embargo, no se incluyen sujetos con gafas de sol, dado que la metodología de detección basada en puntos de referencia faciales (Landmarks) requiere la visibilidad de la región ocular para el cálculo de la fatiga

El dataset incluye aproximadamente 9,120 imágenes, divididas equitativamente en dos categorías:

- **Sujetos Fatigados (Fatigue Subjects):** Esta clase incluye aproximadamente 4,560 imágenes que capturan los signos visuales de somnolencia: cierre ocular parcial (ptosis) o total (microsueños), relajación de la musculatura facial y bostezos. Detectar esta clase con precisión es crítico, ya que corresponde a la condición de riesgo que el sistema debe identificar para emitir alertas preventivas.
- **Sujetos Activos (Active Subjects):** Esta clase comprende aproximadamente 4,560 imágenes que representan conductores despiertos, con ojos completamente abiertos, mirada atenta y postura facial que refleja control cognitivo y motor. Las imágenes incluyen variaciones en la orientación de la cabeza y expresiones naturales, pero mantienen como característica definitoria la alerta visual.

Las imágenes están en formato RGB con resoluciones variables, lo que obliga a estandarizar las dimensiones durante el preprocesamiento. Aunque no hay datos explícitos sobre diversidad étnica y de género en los metadatos públicos, esta variabilidad permite evaluar indirectamente la robustez demográfica del modelo a través de las métricas de explicabilidad.

Codificación numérica de clases: La asignación de etiquetas numéricas se determina automáticamente por el orden alfabético de los directorios durante la carga de datos:

categories = ["Fatigue Subjects", "Active Subjects"] for category in categories: class_num = categories.index(category)

Esta convención resulta en:

- **Fatigue Subjects → Clase 0** (clase negativa en clasificación binaria)
- **Active Subjects → Clase 1** (clase positiva en clasificación binaria)

Implicación metodológica: Dado que el Modelo Base fue obtenido como modelo preentrenado de fuente externa con esta codificación establecida, el Modelo Optimizado desarrollado en este trabajo debió mantener la misma convención para garantizar comparabilidad directa de métricas y resultados. Esta codificación invierte la convención intuitiva común en detección de eventos anómalos, donde típicamente la condición de riesgo (Fatigue) sería la clase positiva. Las implicaciones de esta convención para la interpretación de métricas de clasificación se detallan en la Sección 4.1.

3.2.2 Estrategia de Particionamiento de Datos

Para garantizar resultados válidos y prevenir el sobreajuste, donde el modelo memoriza ejemplos en lugar de aprender patrones generalizables, se implementó un protocolo de división estratificado. La partición de datos quedó configurada de la siguiente manera:

Conjunto de Entrenamiento (80%): Aproximadamente 7,296 imágenes utilizadas exclusivamente para ajustar los pesos del modelo mediante el algoritmo de optimización y retropropagación.

Conjunto de Prueba (20%): Aproximadamente 1,824 imágenes mantenidas completamente aisladas durante todo el entrenamiento e ingeniería de características, empleándose únicamente para la evaluación final del clasificador.

La estratificación asegura que ambas clases (Fatigue y Active) mantengan la proporción 50:50 en los conjuntos de entrenamiento y prueba, preservando el balance del dataset original. El particionamiento se realizó utilizando train_test_split de scikit-learn

con random_state=42 y stratify=labels, garantizando reproducibilidad de los experimentos.

3.3 Fase 3: Preparación de los datos

Para cumplir con los objetivos específicos de auditoría y mejora del modelo, se diseñaron dos flujos de trabajo (*pipelines*) de preparación de datos contrastantes.

3.3.1 Escenario A (Baseline): Inyección de Características Geométricas (Feature Baking)

A diferencia de los enfoques convencionales que procesan imágenes RGB sin modificar, en este escenario se implementó una estrategia de ingeniería de características visuales mediante el uso de MediaPipe Face Mesh. Esta herramienta detecta 468 puntos de referencia distribuidos por toda la región facial, cuyas conexiones forman una malla tridimensional que describe la geometría del rostro.

Pipeline de preprocesamiento

El flujo de procesamiento aplicado a cada imagen consta de cuatro etapas secuenciales:

1. **Detección del rostro**: Se empleó el clasificador Haar Cascade (haarcascade_frontalface_default.xml) con parámetros scaleFactor=1.3 y minNeighbors=5 para localizar y aislar la región facial dentro de la imagen completa. Dado que el dataset contiene únicamente imágenes con un rostro por fotografía, se procesa directamente la ROI detectada mediante las coordenadas del cuadro delimitador (x, y, w, h).
2. **Proyección de la malla facial (Feature Baking)**: Una vez extraída la ROI, se utiliza mp_drawing.draw_landmarks de MediaPipe para superponer digitalmente la teselación completa (FACEMESH_TESSELATION) directamente sobre la imagen RGB. La configuración establece líneas blancas con thickness=1, circle_radius=1 y color= (255, 255, 255), generando una representación híbrida donde las estructuras geométricas clave quedan resaltadas mediante trazos blancos sobre la piel.
3. **Normalización dimensional**: Cada imagen procesada se redimensiona mediante interpolación bilineal a 145×145 píxeles, manteniendo los 3 canales RGB en formato compatible con la arquitectura de la red neuronal.
4. **Escalado de intensidad:** Los valores de píxel se normalizan de [0, 255] al rango [0.0, 1.0] mediante división por 255.0.

El propósito de este enfoque era facilitar la convergencia del modelo al proporcionarle explícitamente la geometría facial relevante para la detección de somnolencia. Sin embargo, esta estrategia planteaba una limitación importante: existía el riesgo de que la CNN aprendiera a identificar las líneas artificiales dibujadas en lugar de las características biométricas reales subyacentes como la textura de la piel o el grado de apertura palpebral, generando así una dependencia de artefactos sintéticos que no estarían presentes en condiciones reales de operación.

3.3.2 Escenario B (Propuesto): Focalización Biométrica y Supresión de Contexto (Propuesta XAI-Guided)

El análisis preliminar de explicabilidad realizado sobre el modelo base (Sección 4.1) reveló una vulnerabilidad crítica: la dispersión de la atención de la red neuronal hacia regiones no causales, como la vestimenta del conductor, el reposacabezas y el fondo. Para rectificar este comportamiento y forzar al modelo a aprender características robustas, se desarrolló un nuevo pipeline de preprocesamiento denominado Focalización Biométrica con Supresión Gaussiana.

Este esquema va más allá de un recorte estático. Se trata de un proceso algorítmico dinámico diseñado para emular la atención selectiva foveal. Su objetivo es preservar la alta frecuencia espacial exclusivamente en las regiones fisiológicamente determinantes para la somnolencia (ojos y boca), tal como dicta el estándar PERCLO [36], mientras se degrada sistemáticamente la información contextual irrelevante para evitar el aprendizaje de correlaciones espurias.

Implementación Algorítmica

El preprocesamiento se ejecuta en cuatro etapas secuenciales para cada fotograma de entrada:

Etapa 1: Mapeo de Topología Facial de Alta Fidelidad

Se utiliza MediaPipe Face Mesh configurado con detección refinada de landmarks (refine_landmarks=True). Este modelo estima la geometría facial tridimensional mediante un grafo de 468 puntos de referencia $L = \{p_1, p_2, \dots, p_{468}\}$, donde cada $p_i = (x, y, z)$representa una coordenada anatómica precisa [65].

La ventaja sobre detectores convencionales como Haar Cascades o HOG radica en la densidad de puntos: en lugar de un cuadro delimitador ruidoso que encierra todo el rostro, se obtienen coordenadas sub-milimétricas de estructuras específicas como párpados, comisuras labiales y contorno mandibular. Esta precisión es fundamental para las etapas posteriores y mantiene robustez ante oclusiones parciales y variaciones de iluminación [70].

Etapa 2: Supresión de Frecuencias Espaciales del Fondo

Para eliminar las texturas del fondo que el modelo podría asociar erróneamente con la clase objetivo (texturas específicas de la tapicería correlacionadas con un sujeto del set de entrenamiento), se aplica un kernel Gaussiano de 99×99 píxeles sobre la totalidad de la imagen original. Matemáticamente, la imagen desenfocada I_{blur} se obtiene mediante:

$$G(u,v) = \frac{1}{2\pi\sigma^2} e^{-\frac{u^2+v^2}{2\sigma^2}}$$

$$I_{blur}(x,y) = \sum_{u=-k}^{k} \sum_{v=-k}^{k} I(x-u, y-v) \times G(u,v)$$

donde $2k+1 = 99$. El tamaño inusualmente grande del kernel no es arbitrario: actúa como filtro paso bajo agresivo que elimina texturas del asiento, patrones de vestimenta y detalles del entorno vehicular, generando una capa base donde solo permanecen formas difusas y bloques de color [71]. Esto previene que la red aprenda asociaciones entre, por ejemplo, un tipo específico de tapicería y la clase "somnoliento" simplemente porque varios sujetos somnolientos en el conjunto de entrenamiento fueron fotografiados en el mismo vehículo.

Etapa 3: Preservación Selectiva de Regiones Críticas

Sobre la imagen desenfocada se restaura la nitidez original únicamente en las zonas biométricamente relevantes. Utilizando los 468 landmarks detectados en la Etapa 1, se construyen dos máscaras binarias rectangulares mediante cv2.boundingRect:

ROI Ocular (M_{ojos}): Definida por los landmarks correspondientes a párpados superior e inferior y cantos del ojo según la topología de MediaPipe Face Mesh (índices: 33, 133, 362, 263, 159, 145, 386, 374, 46, 105, 334, 285). Se aplica un padding de 8

píxeles para incluir la región periorbital donde se manifiesta la ptosis y las cejas, cuya posición delata el esfuerzo por mantener los ojos abiertos [65].

ROI Bucal (M_{boca}): Abarca el contorno labial completo y zona peribucal según los landmarks correspondientes en la topología de MediaPipe (índices: 61, 291, 13, 14, 78, 308, 17, 84, 181, 91, 314, 402, 318, 324, 308, 78). El padding aquí es de 12 píxeles, mayor que en la región ocular, para capturar la apertura máxima de la mandíbula durante bostezos profundos sin que el movimiento exceda los límites de la máscara.

La imagen final se construye mediante composición lineal ponderada por la máscara binaria combinada $M_{total} = M_{ojos} \cup M_{boca}$:

$$I_{final} = (I \odot M_{total}) + (I_{blut} \odot (1 - M_{total}))$$

donde $\odot$ denota el producto de Hadamard (multiplicación elemento a elemento). En términos de operaciones OpenCV, esto se implementa mediante cv2.bitwise_and para extraer las regiones nítidas y cv2.add para combinarlas con el fondo desenfocado. El resultado es una imagen híbrida donde ojos y boca permanecen con toda su información de alta frecuencia (bordes de párpados, textura labial, microexpresiones) mientras que el resto del rostro y el fondo exhiben únicamente formas difusas.

Etapa 4: Recorte Adaptativo y Normalización (Smart Zoom)

Para maximizar la resolución efectiva de las regiones de interés dentro del tensor de entrada fijo, se calcula el bounding box mínimo que encierra todos los 468 landmarks faciales. Las coordenadas extremas $(x_{min}, y_{min}, x_{max}, y_{max})$definen este rectángulo, al cual se añade un margen de seguridad de 30 píxeles en todas direcciones:

$$ROI_{crop} = I_{final}[y_{min} - 30 : y_{max} + 30, x_{min} - 30 : x_{max} + 30]$$

El padding de 30 píxeles previene que movimientos bruscos de la cabeza o bostezos particularmente amplios resulten en cortes inadvertidos del contorno facial. Además, el manejo explícito de bordes evita índices fuera de rango cuando el rostro aparece cerca de los límites del fotograma original.

Este recorte dinámico cumple dos propósitos: primero, incrementa drásticamente la densidad de información biométrica al eliminar píxeles irrelevantes del fondo, permitiendo que ojos y boca ocupen una proporción mayor del tensor de entrada de 145×145 píxeles; segundo, posibilita la detección de estados sutiles de cierre parcial

(ptosis) y microexpresiones que quedarían imperceptibles si el rostro ocupara solo una fracción pequeña de la imagen.

La imagen recortada se redimensiona mediante interpolación bilinear a 145×145 píxeles y se normaliza al rango [0.0, 1.0] dividiéndola por 255.0.

Resultado Final

El pipeline produce imágenes donde el contexto ambiental ha sido deliberadamente degradado mediante supresión de frecuencias espaciales altas (texturas del asiento, volante, ventanas), mientras que los rasgos faciales relevantes permanecen intactos y ocupan una mayor proporción del tensor de entrada gracias al recorte adaptativo. A diferencia del Escenario A, que recurría a máscaras negras artificiales, este enfoque no introduce artefactos gráficos que pudieran funcionar como atajos espurios. La red neuronal ve únicamente características biométricas genuinas en alta resolución sobre un fondo naturalmente degradado, forzándola a basar sus predicciones en los indicadores fisiológicos establecidos en la Sección 2.4.

3.4 Fase 4: Modelado y Arquitectura de la Red

Para permitir una comparación controlada entre escenarios, se mantuvo una arquitectura de red neuronal convolucional similares, pero no idénticas, presentando diferencias específicas en configuración de padding y regularización.

3.4.1 Estructura Compartida

Ambos modelos utilizan una arquitectura convolucional secuencial que consta de:

- **Bloque Convolucional 1:**
  - Conv2D: 16 filtros (3×3), activación ReLU
  - BatchNormalization
  - MaxPooling2D (2×2)
- **2. Bloque Convolucional 2:**
  - Conv2D: 32 filtros (5×5), activación ReLU
  - BatchNormalization
  - MaxPooling2D (2×2)
- **3. Bloque Convolucional 3:**
  - Conv2D: 64 filtros (10×10), activación ReLU
  - BatchNormalization

    - MaxPooling2D (2×2)
- **4. Bloque Convolucional 4 (Target Layer):**
    - Conv2D: 128 filtros (12×12), activación ReLU
    - BatchNormalization
- **5. Clasificador Denso:**
    - Flatten
    - Dense (128 neuronas, ReLU)
    - Dropout para regularización
    - Dense (64 neuronas, ReLU)
    - Dense (1 neurona, Sigmoid) para clasificación binaria

3.4.2 Diferencias arquitectónicas clave

Tabla 5. Diferencias arquitectónicas entre configuraciones del modelo CNN

Aspecto	Modelo Base	Modelo Optimizado
Padding convolucional	`'valid'` (por defecto)	`'same'` (explícito)
Dropou trate	0.25	0.2
Nombre capa objetivo	Sin nombre explícito	`name='target_conv'`
Optimizador	`optimizer="adam"` (lr=0.001 por defecto)	`Adam(learning_rate=0.001)` explícito

La diferencia más significativa radica en la estrategia de padding: el Modelo Base utiliza padding='valid' por defecto, lo que provoca que las capas convolucionales reduzcan las dimensiones espaciales en cada operación, generando una reducción acumulada considerable con kernels de tamaños 3, 5, 10 y 12 antes de las operaciones de pooling. Por el contrario, el Modelo Optimizado emplea padding='same', preservando las dimensiones espaciales después de cada convolución y permitiendo que solo las operaciones de MaxPooling controlen la reducción dimensional, resultando en mapas de características ligeramente más grandes en capas intermedias. Esta diferencia puede afectar la capacidad del modelo para capturar información espacial de alta resolución, especialmente relevante para detectar características sutiles como el grado de apertura palpebral.

3.5 Fase 5: Evaluación

3.5.1 Estrategia de Selección de Casos de Estudio

Para validar la interpretabilidad de los modelos más allá de su precisión global, se diseñó un protocolo de evaluación cualitativa mediante muestreo dirigido. Esta fase busca auditar la coherencia fisiológica de las decisiones de la red, más que evaluar su rendimiento estadístico (detallado en el Capítulo 4).

Es fundamental mencionar que la selección de estas muestras específicas se realizó tomando como referencia las predicciones generadas por el Modelo Base (Escenario A) con el objetivo de aislar cinco casos críticos del conjunto de prueba mediante muestreo no probabilístico intencional, basado en tres criterios:

1. **Verificación de aciertos (VP y VN):** Muestras correctamente clasificadas en ambas clases ("Activo" y "Fatiga"). El objetivo fue confirmar que las predicciones acertadas se fundamentan en características faciales relevantes (como apertura ocular) y descartar el aprendizaje de patrones espurios o elementos irrelevantes del fondo (efecto Clever Hans).
2. **Análisis de errores (FP y FN):** Casos con predicciones incorrectas. El propósito fue examinar las visualizaciones para identificar los elementos que confundieron a la red, ya sean oclusiones, condiciones atípicas de iluminación o gestos ambiguos (bostezos, sonrisas) interpretados erróneamente.
3. **Análisis de incertidumbre (casos ambiguos):** Muestras donde la red mostró baja confianza en su predicción (probabilidad ≈ 0.50). Se evaluó cómo las técnicas XAI visualizan la distribución de pesos positivos y negativos ante evidencia contradictoria.

3.5.2 Configuración Experimental de Técnicas XAI

Una vez definidos los escenarios de preprocesamiento y seleccionados los casos de estudio, se procedió a configurar las cuatro técnicas de explicabilidad que permitirían analizar el comportamiento del modelo en cada contexto. La configuración de estos métodos requirió ajustes específicos para adaptarse a las particularidades de cada escenario, manteniendo la comparabilidad entre ambos.

3.5.2.1 Grad-CAM

Selección de capa objetivo

La capa convolucional analizada difiere entre escenarios:

- **Escenario A:** conv2d_2 (64 filtros, kernel 10×10). Aunque el modelo posee una capa más profunda (conv2d_3), se determinó empíricamente que conv2d_2 preserva mejor la información espacial de los rasgos faciales antes de la reducción excesiva por el pooling.
- **Escenario B:** Última capa Conv2D identificada dinámicamente mediante búsqueda inversa en la arquitectura del modelo (target_conv).

Parametros de implementación

Tabla 6. Parámetros de configuración de Grad-CAM para análisis de explicabilidad

Parámetro	Valor
Número de filtros	64 modelo Base, 128 modelo Nuevo
Estrategia de gradiente	∂Score/∂Activation
Pooling de gradientes	Global Average
Normalización	ReLU + Min-Max
Colormap	JET
Superposición	60% img + 40% heatmap

3.5.2.2 Score-CAM

Se incorporó Score-CAM como alternativa libre de gradientes (gradient-free) para validar los hallazgos de Grad-CAM, evitando problemas de gradientes ruidosos o saturados. Ambos escenarios utilizan la misma capa de extracción que Grad-CAM, garantizando comparabilidad directa.

Tabla 7. Parámetros de configuración de Score-CAM para análisis libre de gradientes

Parámetro	Valor
Número de máscaras	Dinámico (todos los canales de la capa objetivo)
Redimensionamiento	Bilinear → 145×145
Normalización por canal	Min-Max individual

- **Proceso:**

1. Cada mapa de activación se redimensionó a 145×145 píxeles y se normalizó al rango [0, 1].
2. Se generó un producto de Hadamard entre el mapa normalizado y la imagen original.
3. Las imágenes enmascaradas se reintrodujeron al modelo para obtener el puntaje de confianza (CIC - Channel Importance Confidence).
4. El mapa final es una combinación lineal ponderada de las activaciones que más contribuyeron a la predicción final.

3.5.2.3 LIME

Para garantizar consistencia metodológica y comparabilidad directa entre escenarios, se unificó la estrategia de segmentación utilizando el algoritmo SLIC (Simple Linear Iterative Clustering) con parámetros idénticos en ambos casos.

Configuración unificada con SLIC

Tabla 8. Parámetros de segmentación y perturbación para LIME

Parámetro	Valor
Algoritmo	SLIC
n_segments	150 (forzado)
compactness	10
sigma	1
Tamaño promedio	15-30 px^2
Muestras de perturbación	1,000
Features mostradas	Top 15
Modelo local	Ridge Regression (por defecto)
Estrategia de ocultación	hide_rest=False

La segmentación fina con 150 superpíxeles permite capturar detalles locales tanto en las características globales del rostro (Escenario A) como en las regiones críticas delimitadas por la máscara de precisión (Escenario B). Los superpíxeles de menor tamaño facilitan la evaluación de contribuciones específicas en zonas como la posición de párpados o apertura bucal.

3.5.2.4 SHAP

Conjunto de fondo (Background Dataset)

El conjunto de referencia se construyó mediante muestreo aleatorio del dataset principal, aplicando el preprocesamiento correspondiente a cada escenario:

Tabla 9.Características del conjunto de fondo (background dataset) para SHAP

Característica	Escenario A	Escenario B
Tamaño	100 imágenes	100 imágenes
Selección	Muestreo aleatorio	Muestreo aleatorio
Preprocesamiento	Pipeline FaceMesh completo	Pipeline máscara de precisión
Justificación	Primeras 100 válidas	Primeras 100 válidas

Ambos escenarios utilizan el mismo tamaño de conjunto de fondo para mantener comparabilidad en el cálculo de valores SHAP, aunque el preprocesamiento difiere según el pipeline correspondiente.

Parámetros de cálculo

Tabla 10. Parámetros de cálculo y visualización de valores SHAP

Parámetro	Valor
Tipo de explicador	GradientExplainer
Casos evaluados	1 imagen por ejecución
Formato de salida	(1, 145, 145, 3)
Normalización de heatmap	Percentil 99.9
Colormap	Seismic (divergente) rojo contribuye y azul va en contra.
Transparencia	Basada en magnitud absoluta

Proceso de visualización:

1. Los valores SHAP se calculan para cada píxel en los tres canales de color.

2. Se genera un mapa de calor agregado sumando las contribuciones a través de los canales.
3. La normalización se realiza usando el percentil 99.9 de valores absolutos para evitar outliers (valores atípicos extremos).
4. El mapa final se superpone sobre la imagen en escala de grises con transparencia proporcional a la magnitud de contribución.

3.6 Métricas de Evaluación de Explicabilidad

La validación cuantitativa se llevó a cabo mediante tres métricas estándar. Para asegurar la objetividad de los resultados, la evaluación se realizó con el conjunto de prueba (Test Set), el cual contiene datos que permanecieron sin uso durante el entrenamiento. A partir de este conjunto, se seleccionó de manera aleatoria una muestra representativa correspondiente al 10% del total ($n \approx 182$ imágenes). La selección se realizó aplicando un muestreo estratificado que permitió mantener el equilibrio entre las diferentes clases.

Tabla 11. Parámetros de configuración de las métricas de explicabilidad

Parámetro	Valor	Configuración
Tamaño de Entrada	145 × 145 píxeles	RGB (3 canales)
División de Datos	80% / 20%	Entrenamiento / Prueba
Muestra XAI	10% del Test	n ≈ 182 imágenes
Estrategia	Muestreo Estratificado	random_state = 42
Técnica XAI	Grad-CAM	Gradient-weighted CAM
Librerías	TensorFlow 2.x	OpenCV, MediaPipe, Scikit-learn

3.6.1 Cálculo de la Fidelidad (Faithfulness)

Para implementar esta métrica se utilizó una técnica de oclusión basada en umbrales (Threshold-based Occlusion), donde el algoritmo identifica los píxeles de mayor intensidad en el mapa Grad-CAM, específicamente aquellos ubicados en el percentil 80, lo que representa el 20% superior de valores. Con estos píxeles se genera una máscara binaria que permite anular las regiones correspondientes en la imagen original mediante la asignación del valor 0. Posteriormente, la métrica se calcula

determinando la diferencia absoluta entre la probabilidad de predicción inicial (P_{orig}) y la probabilidad obtenida tras aplicar la perturbación (P_{occ}):

$$Fidelidad = \left| P_{orig} - P_{ocult} \right|$$

Un valor alto indica que ocultar las zonas resaltadas impactó fuertemente en la decisión del modelo, lo que confirma la veracidad de la explicación proporcionada por la técnica XAI.

3.6.2 Cálculo de la Robustez (Robustness)

Para aplicar esta métrica, se incorporó ruido Gaussiano con parámetros $\mu = 0$y $\sigma = 0.05$ a cada imagen del conjunto de prueba normalizado, simulando pequeñas perturbaciones sensoriales. A partir de estas imágenes alteradas se generaron nuevos mapas de calor, los cuales fueron comparados con los mapas originales mediante el cálculo del Índice de Similitud Estructural (SSIM) entre el mapa de la imagen sin alteraciones (H_{orig}) y el mapa correspondiente a la versión con ruido (H_{noise}):

$$Robustez = SSIM(H_{orig}, H_{noise})$$

Los valores cercanos a 1 indican que la técnica de explicación mantiene su coherencia estructural ante perturbaciones imperceptibles, lo cual resulta fundamental para aplicaciones en tiempo real donde las condiciones de captura pueden fluctuar ligeramente.

3.6.3 Cálculo de la Localización (Energy Pointing Game)

Se automatizó la métrica de Pointing Game para verificar si la atención del modelo recae anatómicamente en los ojos, región fisiológicamente relevante para la detección de somnolencia. En lugar de depender de anotaciones manuales, se emplearon los puntos de referencia (landmarks) faciales de MediaPipe para generar dinámicamente una máscara binaria precisa (M_{ojos}) que delimita exclusivamente las regiones del ojo izquierdo y derecho, garantizando consistencia y eliminando la subjetividad de la anotación humana. La métrica se calculó como la proporción de energía acumulada del mapa de calor dentro de dicha máscara en relación con la energía total de la imagen:

$$Localización = \frac{\sum (x, y) \in M_{ojos}\, H(x, y)}{\sum (x, y) \in Imagen\ \ H(x, y)}$$

Valores elevados de esta métrica indican que la técnica concentra efectivamente la explicación en las áreas anatómicamente correctas, validando que el modelo aprende patrones fisiológicos correctos en lugar de correlaciones erróneas con otros elementos de la imagen.

CAPÍTULO IV

4. RESULTADOS Y ANÁLISIS

Este capítulo presenta los resultados obtenidos tras evaluar el Modelo Base y el Modelo Optimizado desarrollado. La validación se estructura en dos dimensiones complementarias: primero, un análisis estadístico del rendimiento de clasificación y, segundo, una evaluación cualitativa y cuantitativa de la interpretabilidad mediante técnicas de Inteligencia Artificial Explicable (XAI). Cabe destacar que, con el fin de garantizar la transparencia de estos resultados, todas las métricas presentadas derivan de una evaluación dinámica ejecutada de forma autónoma. Se descartaron las métricas de referencia del repositorio original (Kaggle) tras detectar que estas no eran reproducibles debido a inconsistencias en el script base, estableciendo así una línea base real y verificable para esta investigación.

4.1 Resultados del Desempeño de Clasificación

Esta sección compara el Modelo Base (Escenario A: Feature Baking) con el Modelo Optimizado (Escenario B: Máscara Selectiva con Zoom Adaptativo). El objetivo es evaluar la capacidad de ambos modelos para distinguir entre estados de "Fatiga" y "Activo" en conductores.

Nota sobre convenciones de etiquetado: Según lo detallado en la Sección 3.2.1, la codificación de clases está determinada por el orden alfabético de los directorios del dataset:

- **Fatigue Subjects → Clase 0** (clase negativa)
- **Active Subjects → Clase 1** (clase positiva)

Esta convención, heredada del Modelo Base preentrenado, invierte la intuición habitual donde el evento de riesgo (fatiga) sería la clase positiva. Esta inversión tiene implicaciones directas en la interpretación de errores, especialmente al distinguir entre falsos positivos y falsos negativos.

4.1.1 Matrices de Confusión

Las matrices de confusión permiten visualizar no solo los aciertos del algoritmo, sino también la distribución y naturaleza de los errores cometidos.

Convención de lectura:

Las filas representan la clase real (ground truth) y las columnas la predicción del modelo. Las celdas se interpretan según el diagrama siguiente:

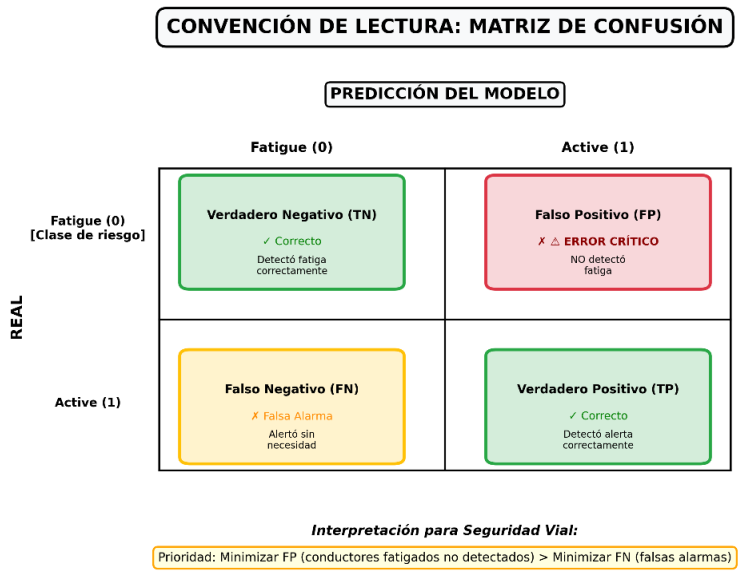

Figura 3. Convención de lectura para las matrices de confusión

Esta figura ilustra la estructura y convención de lectura empleada para interpretar las matrices de confusión presentadas en este capítulo. El eje vertical (REAL) representa la clase verdadera del conductor según el ground truth del dataset, mientras que el eje horizontal (PREDICCIÓN DEL MODELO) indica la clasificación generada por el algoritmo. Las cuatro celdas de la matriz corresponden a:

- **Verdadero Negativo (TN)** [verde, superior izquierda]: Conductores en estado de fatiga correctamente identificados como "Fatiga". Representa detección exitosa del riesgo. - **Falso Positivo (FP)** [rojo, superior derecha]: Conductores fatigados clasificados erróneamente como "Activo". Este error es crítico para la seguridad vial, ya que representa casos de fatiga no detectados.
- **Falso Negativo (FN)** [amarillo, inferior izquierda]: Conductores activos clasificados incorrectamente como "Fatiga". Este error genera falsas alarmas que pueden afectar la confianza del usuario en el sistema.
- **Verdadero Positivo (TP)** [verde, inferior derecha]: Conductores en estado activo correctamente identificados como "Activo". Confirma el funcionamiento adecuado del sistema ante conductores alertas.

Prioridad para seguridad vial: Desde la perspectiva de seguridad, el objetivo prioritario es minimizar los Falsos Positivos (conductores fatigados no detectados), incluso si esto implica un ligero incremento en Falsos Negativos (falsas alarmas), dado que las consecuencias de no detectar fatiga son potencialmente fatales.

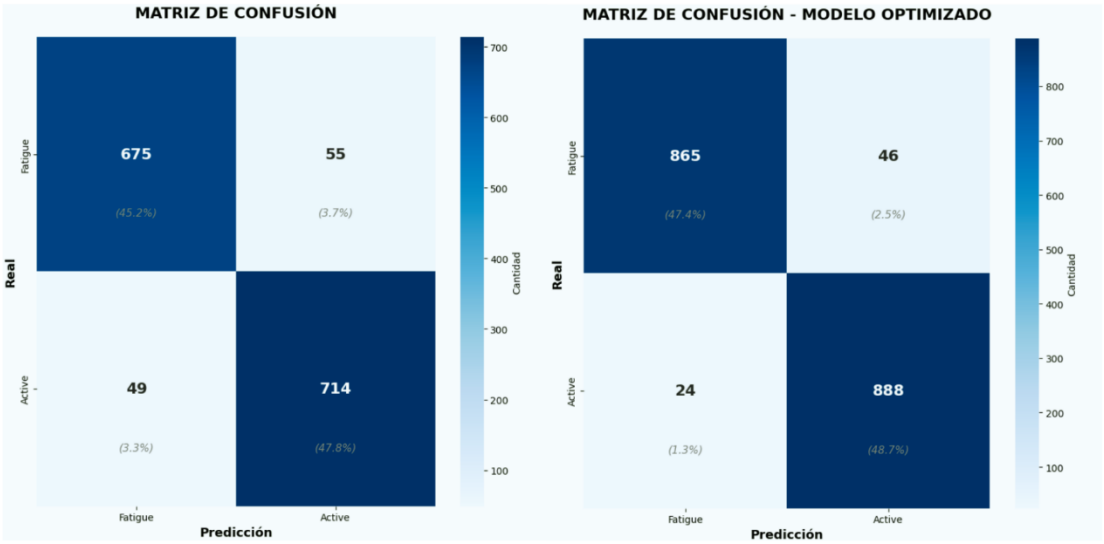

Figura 4. Comparación de Matrices de Confusión: Modelo Base (izquierda) y Modelo Optimizado (derecha)

El análisis comparativo de ambos modelos revela mejoras sustanciales en todas las dimensiones evaluadas. En cuanto a la tasa de procesamiento, el modelo base logró procesar exitosamente 1,493 imágenes de las 1,824 disponibles (81.85%), perdiendo 331 imágenes debido a fallos en la detección facial mediante Haar Cascade, lo que evidenció limitaciones importantes en la robustez del preprocesamiento ante condiciones desafiantes como variaciones de iluminación, ángulos faciales no frontales y oclusiones parciales. Por su parte, el modelo optimizado procesó 1,823 imágenes (99.95%), incrementando en 330 imágenes la capacidad de procesamiento gracias a la implementación de MediaPipe Face Mesh, que demostró mayor robustez ante condiciones adversas.

En términos de exactitud global, el modelo base alcanzó un 93.03%, identificando correctamente 675 casos de fatiga y 714 casos activos. Sin embargo, presentó 55 falsos positivos (conductores fatigados clasificados erróneamente como activos, representando un riesgo significativo para la seguridad vial) y 49 falsos negativos (conductores activos

clasificados como fatigados, generando falsas alarmas). El modelo optimizado mejoró la exactitud a 96.16%, reduciendo los falsos positivos a 46 casos (-16.4%) y disminuyendo drásticamente los falsos negativos a 24 casos (-51.0%), lo que representa una mejora crítica tanto en seguridad vial como en la confiabilidad del sistema de alertas.

4.1.2 Métricas Globales de Clasificación

A partir de las matrices anteriores, se extrajeron métricas comparativas que resumen el desempeño global de ambos modelos. Estas métricas fueron calculadas.

Tabla 12. Comparación de métricas globales de rendimiento entre Modelo Base y Modelo Optimizado.

Métrica	Modelo Base	Modelo Optimizado	Incremento (puntos porcentuales)	Mejora relativa
Exactitud (Accuracy)	93.03%	96.16%	+3.13	+3.36%
Precisión	92.85%	95.07%	+2.22	+2.39%
Sensibilidad (Recall)	93.58%	97.37%	+3.79	+4.05%
F1-Score	93.21%	96.21%	+3.00	+3.22%

Estas métricas consideran la clase "Active" como positiva, dado que el sistema debe maximizar la detección de conductores activos versus fatigados para evaluar la sensibilidad en la identificación de estados de alerta.

4.1.3 Análisis por Clases

El desempeño detallado para ambas clases revela diferencias importantes en cómo cada modelo procesa las dos categorías de estado del conductor.

Modelo Base

Tabla 13. Métricas de clasificación por clase para el Modelo Base.

Clase	Precisión	Sensibilidad (Recall)	F1-Score	Soporte

Fatigue	93.23%	92.47%	92.85%	730
Active	92.85%	93.58%	93.21%	763

Modelo Optimizado

Tabla 14. Métricas de clasificación por clase para el Modelo Optimizado.

Clase	Precisión	Sensibilidad (Recall)	F1-Score	Soporte
Fatigue	97.30%	94.95%	96.11%	911
Active	95.07%	97.37%	96.21%	912

Análisis por clase crítica (Fatigue)

Dado que la detección de fatiga es el objetivo principal del sistema, resulta pertinente examinar en detalle el desempeño en esta clase de riesgo.

Tabla 15. Análisis comparativo de la clase crítica (Fatigue) entre ambos modelos.

Métrica	Modelo Base	Modelo Optimizado	Mejora
Precisión (Fatigue)	93.23%	97.30%	+4.07%
Sensibilidad (Fatigue)	92.47%	94.95%	+2.48%
Tasa de Falsos Positivos (FPR)	7.53%	5.05%	-2.48%
Tasa de Falsos Negativos (FNR)	6.42%	2.63%	-3.79%

4.1.4 Análisis de Errores

La caracterización de los errores permite evaluar el impacto operativo de cada tipo de fallo en la seguridad del sistema.

Errores críticos del Modelo Base:

- **Falsos Negativos:** 49 casos (6.42%) - Conductores activos clasificados erróneamente como fatigados (falsas alarmas)
- **Falsos Positivos:** 55 casos (7.53%) - Conductores fatigados NO detectados (riesgo crítico de seguridad)

Errores críticos del Modelo Optimizado**:**

- **Falsos Negativos:** 24 casos (2.63%) - Reducción de **25 casos** (-51.0%)
- **Falsos Positivos:** 46 casos (5.05%) - Reducción de **9 casos** (-16.4%)

4.1.5 Síntesis de los resultados

El análisis comparativo evidencia mejoras significativas del Modelo Optimizado en todas las métricas evaluadas. La reducción de falsos positivos de 55 a 46 casos (-16.4%) y de falsos negativos de 49 a 24 casos (-51.0%) representa avances críticos tanto para la seguridad como para la usabilidad del sistema. El incremento en sensibilidad de 93.58% a 97.37% (+3.79 puntos porcentuales) confirma una mayor capacidad de detección de estados de alerta.

Estas mejoras tienen su origen en la estrategia de máscara selectiva implementada en el Modelo Optimizado. Esta técnica elimina sistemáticamente el ruido contextual (elementos del vehículo, fondo, iluminación ambiental) mediante desenfoque gaussiano del contexto no relevante, mientras preserva de forma nítida las regiones anatómicas críticas: la zona periocular y bucal. Este enfoque aplica los principios de simulación de visión foveal y supresión de fondo, alineando la atención del modelo con los parámetros biométricos del estándar PERCLOS descritos en la sección 2.5.

Por otro lado, el incremento en la tasa de procesamiento de 81.85% a 99.95% (+330 imágenes adicionales) demuestra la superior robustez de MediaPipe Face Mesh frente a Haar Cascade en condiciones desafiantes, lo que constituye una mejora fundamental en la viabilidad práctica del sistema.

Habiendo establecido las mejoras cuantitativas en rendimiento de clasificación, la siguiente sección analiza cualitativamente cómo el modelo optimizado logró estos resultados mediante el examen de sus patrones de atención utilizando técnicas de Inteligencia Artificial Explicable.

4.2. Análisis Cualitativo mediante Técnicas de Inteligencia Artificial Explicable (XAI)

En esta sección se contrastan las interpretaciones visuales generadas por las técnicas XAI descritas en las Secciones 2.1.5-2.1.8 (Grad-CAM, Score-CAM, LIME y SHAP) para el Modelo Base frente al Modelo Optimizado. El objetivo principal es demostrar cómo las técnicas de explicabilidad pueden identificar falencias y sesgos en la arquitectura inicial del modelo, permitiendo realizar ajustes fundamentados en el preprocesamiento de datos.

A partir del análisis XAI del Modelo Base, se detectaron patrones de atención dispersos en regiones anatómicamente irrelevantes: fondo, vestimenta, elementos del vehículo. Este comportamiento evidencia el fenómeno "Clever Hans" descrito en la Sección 2.2.2, donde los modelos aprenden correlaciones espurias en lugar de características causales. Este hallazgo motivó el desarrollo de la estrategia de máscara selectiva implementada en el Modelo Optimizado, siguiendo la metodología de Ingeniería de Datos Guiada por XAI (Sección 2.3).

La comparación posterior mediante las mismas técnicas XAI permite validar que las mejoras implementadas lograron efectivamente alinear la atención del modelo con regiones fisiológicamente relevantes (zona periocular y bucal), comprobando que los ajustes basados en interpretabilidad resultaron en predicciones más robustas y coherentes con criterios médicos de detección de somnolencia.

4.2.1 Caso 1: Verdadero Positivo (Conductor Activo)

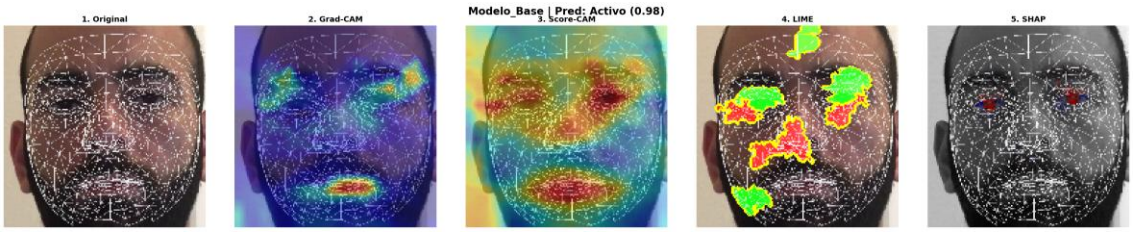

Figura 5. Análisis XAI – Caso 1 (Modelo Base)

Análisis del Modelo Base:

Predicción: Activo (confianza: 0.98)

- **Grad-CAM:** Muestra que la red neuronal identifica correctamente la región facial como prioritaria, activándose intensamente en la zona del ojo izquierdo y la mejilla. Sin embargo, revela una limitación importante de robustez al mostrar una atención asimétrica, ignorando casi por completo el ojo derecho. Esto sugiere que la predicción depende excesivamente de la iluminación lateral más que de una visión binocular completa.
- **SCORE-CAM:** Ofrece una visión más estructural, destacando el triángulo central del rostro (nariz y boca) como la zona de mayor importancia para confirmar la clase "Activo". Si bien esto valida la presencia del sujeto, también indica un posible "atajo" en el aprendizaje, ya que el modelo prioriza la estructura de la nariz

y la textura de la piel sobre los ojos mismos. Este comportamiento podría generar falsos positivos si el sujeto tuviera los ojos cerrados, pero mantuviera la cabeza recta.

- **LIME:** Valida parcialmente la lógica de la decisión al marcar las regiones de los pómulos y la zona ocular en verde, confirmando que la textura de la piel y la apertura de los ojos contribuyen positivamente a la predicción. No obstante, la inclusión de grandes áreas de las mejillas sugiere que el modelo no discrimina finamente solo la apertura ocular, sino que se apoya en el contexto general del rostro iluminado.
- **SHAP:** Es la técnica que demuestra la mayor precisión granular en este caso, concentrando los puntos de decisión (rojos y azules) directamente sobre las pupilas y los puntos de referencia de la malla facial. Esto confirma que, a nivel de píxel, el modelo es capaz de detectar la mirada, aunque la dispersión de puntos en el resto de la cara indica que todavía procesa cierto nivel de ruido visual innecesario.

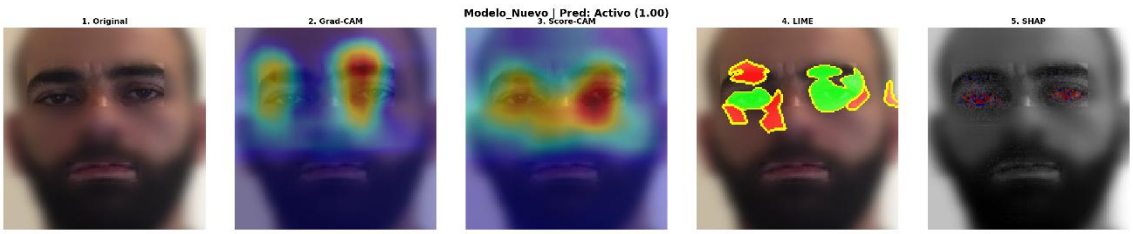

Figura 6. Análisis XAI – Caso 1 (Modelo Optimizado)

Análisis del Modelo Optimizado:

Predicción: Activo (confianza: 1.00).

- **Grad-CAM:** Lo primero que resalta es la simetría perfecta. A diferencia del modelo base, que actuaba como un "cíclope" fijándose solo en un ojo, este modelo optimizado presenta dos manchas rojas intensas y bien definidas sobre ambos ojos simultáneamente. Esto representa un avance técnico considerable en robustez, indicando que el modelo no depende de la iluminación lateral, sino que busca activamente la validación binocular. Ha dejado de buscar señales en las mejillas o la frente para centrarse exclusivamente en la región de interés real.
- **SCORE-CAM:** Este mapa confirma que se ha eliminado el problema de dispersión de atención. La activación (zonas rojas) está confinada estrictamente a

la región de los ojos y el puente nasal, sin fugas hacia el fondo, la ropa o la barbilla. Mientras el modelo base se distraía con ruido del entorno cuando tenía dudas, este modelo demuestra una focalización precisa. El hecho de que el resto de la cara aparezca difuminado o en azul frío sugiere que el modelo ignora deliberadamente la información irrelevante, lo que justifica esa confianza del 100% (1.00)

- **LIME:** Aquí se aprecia una precisión quirúrgica. Los superpíxeles verdes (Apoyan la predicción "Activo") están dibujados casi exclusivamente sobre los globos oculares y los párpados superiores. No hay grandes manchas verdes en los pómulos o la nariz como en el modelo anterior. Esto demuestra que la red neuronal ha aprendido que la característica definitoria de "estar activo" no es simplemente tener cara visible, sino específicamente tener los ojos abiertos.
- **SHAP:** El gráfico es notablemente más limpio y menos ruidoso. Los puntos de decisión (rojos y azules) están densamente empaquetados en las pupilas e iris, con muy poca dispersión en la piel circundante. Esto indica que el modelo es altamente eficiente: necesita procesar menos píxeles para llegar a una conclusión más fuerte.

Síntesis Comparativa: Mientras que el Modelo Base lograba una predicción correcta (0.98) pero técnicamente deficiente (mostrando atención asimétrica sobre un solo ojo y "contaminando" su decisión con ruido de la nariz y las mejillas), el Modelo Optimizado alcanza certeza absoluta (1.00) mediante una corrección estructural total. Sus mapas Grad-CAM exhiben ahora simetría binocular perfecta, y las activaciones de Score-CAM y LIME se han limpiado de todo ruido externo. Esto demuestra que el algoritmo ha dejado de depender de la estructura facial general para especializarse quirúrgicamente en la región orbital, eliminando la vulnerabilidad a la iluminación lateral que caracterizaba a su predecesor.

4.2.2 Caso 2: Verdadero Negativo (Conductor Fatigado)

Contexto: Sujeto con ojos cerrados usando gafas.

Figura 7. Análisis XAI – Caso 2 (Modelo Base)

Análisis del Modelo Base:

Predicción: Fatiga (confianza: 1.00 - probabilidad de Activo: 0.00)

- **Grad-CAM:** Evidencia que el modelo ha aprendido correctamente que el cierre ocular es la señal crítica de fatiga, mostrando un punto caliente rojo intenso sobre el ojo izquierdo cerrado. Sin embargo, al igual que en el caso anterior, la atención es monocular y se centra solo en un lado de la cara. Esto demuestra que el algoritmo no verifica la congruencia de ambos ojos, haciéndolo vulnerable si el ojo vigilado se oculta por sombras o reflejos.
- **SCORE-CAM:** Revela la debilidad más notable en este acierto: las zonas de mayor relevancia no están en los ojos, sino dispersas en la barbilla y la frente. Esto indica que el modelo llegó a la respuesta correcta ("Fatiga") basándose en características secundarias como la relajación muscular o la postura de la cabeza, en lugar de observar directamente la evidencia principal (los párpados). Este comportamiento representa un acierto circunstancial poco robusto.
- **LIME:** Este análisis revela un conflicto interno en el modelo: las áreas verdes (que apoyan "Fatiga") cubren los ojos y las gafas, confirmando que los párpados cerrados constituyen evidencia a favor. Sin embargo, la zona de la boca está marcada en rojo/rosa, indicando que la boca (quizás por estar ligeramente abierta) actúa como evidencia en contra (hacia "Activo"). El modelo predice fatiga porque el peso de los ojos (verde) supera al de la boca (rojo).
- **SHAP:** Proporciona la validación técnica más sólida, delineando con alta precisión la línea de unión de los párpados cerrados mediante una densa nube de puntos azules. Esto demuestra que, matemáticamente, el modelo sí distingue la geometría del ojo cerrado, aunque depende críticamente de que esos píxeles específicos sean visibles y no estén obstruidos por marcos de gafas gruesos o mala iluminación.

Figura 8. Análisis XAI – Caso 2 (Modelo Optimizado)

Análisis del Modelo Optimizado:

Predicción: Fatiga (confianza: 1.00 - probabilidad de Activo: 0.00)

- **Grad-CAM:** En este caso, el mapa de calor muestra una mejora sustancial respecto al modelo base, aunque no es perfecto. La activación principal (zonas rojas y naranjas) se concentra correctamente en la mitad superior del rostro, abarcando la región de los ojos y las sienes. A diferencia del modelo antiguo que solo observaba un ojo, aquí hay actividad bilateral, aunque se nota cierta dispersión hacia los marcos de las gafas y el borde del rostro. Esto sugiere que el modelo utiliza el contraste de las gafas oscuras y la línea de los ojos cerrados como un conjunto de características, en lugar de aislar únicamente los párpados.
- **SCORE-CAM:** Esta técnica revela la eficacia del preprocesamiento (el desenfoque del fondo). La atención del modelo se mantiene estrictamente dentro de la "máscara" clara del rostro, ignorando totalmente la ropa o el fondo que confundían al modelo anterior. Las zonas rojas más intensas se ubican en las esquinas de los ojos y la frente, indicando que el modelo evalúa la falta de tensión en la zona orbital y la frente para confirmar la fatiga. Esto demuestra que no se distrae, aunque su foco es más estructural que puntual.
- **LIME:** El análisis de segmentación es revelador y anatómicamente correcto. Las áreas verdes (que apoyan la predicción de "Fatiga") marcan explícitamente los dos ojos cerrados y, crucialmente, la boca entreabierta. Esto representa un avance importante: el modelo ha aprendido a correlacionar dos señales biológicas distintas (ojos cerrados + boca relajada) para emitir su juicio, mostrando una lógica interna mucho más robusta que la del modelo base, que a menudo entraba en conflicto entre estas dos zonas.
- **SHAP:** Nuevamente, SHAP ofrece la evidencia más fuerte de la precisión del modelo. A pesar de la presencia de gafas (que solían engañar al modelo anterior), la nube de puntos azules se concentra con extrema precisión en la línea de las pestañas y la unión de los párpados. El modelo ignora los reflejos de los lentes y penaliza la probabilidad de "Activo" basándose en los píxeles exactos donde debería estar la pupila y no está. Esto confirma que el modelo Optimizado es capaz de "ver a través" de los accesorios y encontrar la característica de cierre ocular.

Síntesis Comparativa

Mientras que el Modelo Base acertaba la fatiga (0.00) de manera frágil —apoyándose excesivamente en un solo ojo y desviando su atención hacia la barbilla o el fondo para buscar confirmación—, el modelo optimizado demuestra robustez superior gracias a su enfoque selectivo. Grad-CAM y LIME confirman que el algoritmo ahora integra información de ambos ojos y la boca simultáneamente, eliminando las contradicciones internas. Lo más importante es que SHAP prueba que el sistema ha aprendido a ignorar la interferencia de las gafas para detectar el cierre ocular con precisión milimétrica, algo en lo que su predecesor fallaba frecuentemente.

4.2.3 Caso 3: Falso Positivo (Fatiga No Detectada)

Contexto: Sujeto con párpados caídos (mirada "pesada") sugiriendo somnolencia.

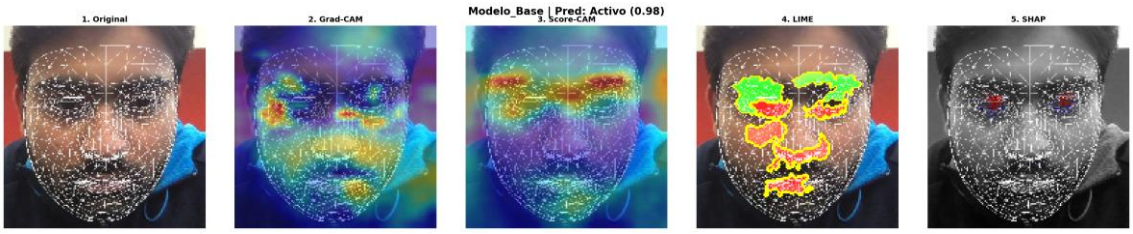

Figura 9. Análisis XAI – Caso 3 (Modelo Base)

Análisis del Modelo Base:

Predicción: Activo (confianza: 0.98) - ERROR CRÍTICO

- **Grad-CAM:** El mapa de calor aquí es disperso y confuso. Las activaciones más fuertes (zonas rojas) no están centradas en las pupilas, sino en los pómulos y las comisuras externas de los ojos. Esto sugiere que el modelo no encontró una señal clara de "ojos cerrados" y se distrajo con el contraste de la piel o las sombras en las mejillas, asumiendo erróneamente actividad por falta de evidencia contundente de cierre ocular.
- **SCORE-CAM:** Esta técnica revela la causa principal del error: el foco de atención está intensamente concentrado en la frente y las cejas (mancha roja superior). Es probable que el conductor, al luchar contra el sueño, haya levantado las cejas o arrugado la frente. El modelo interpretó erróneamente esta tensión muscular en la parte superior del rostro como un signo de "estado de alerta", ignorando la pesadez de los párpados.
- **LIME:** La segmentación muestra una división clara: la zona de los ojos y la frente está marcada en verde (apoya "Activo"), mientras que la nariz y la boca están en

rojo (apoyan "Fatiga"). El modelo dio más peso a la región superior. Esto confirma que, para la red neuronal, la apariencia de los ojos (posiblemente por brillo o forma) y la frente se asemejaban más a patrones de un conductor despierto, superando las señales de fatiga del resto de la cara.

- **SHAP:** Los valores SHAP muestran puntos rojos (que incrementan la probabilidad de "Activo") acumulados directamente sobre las pupilas. Esto es crítico: indica que el modelo analizó los píxeles de los ojos y, a nivel matemático, "creyó" ver ojos abiertos. Esto pudo deberse a un reflejo en la córnea o a que los ojos no estaban completamente cerrados, lo que engañó al algoritmo haciéndole sumar puntos hacia la clase "Activo" en lugar de restarlos.

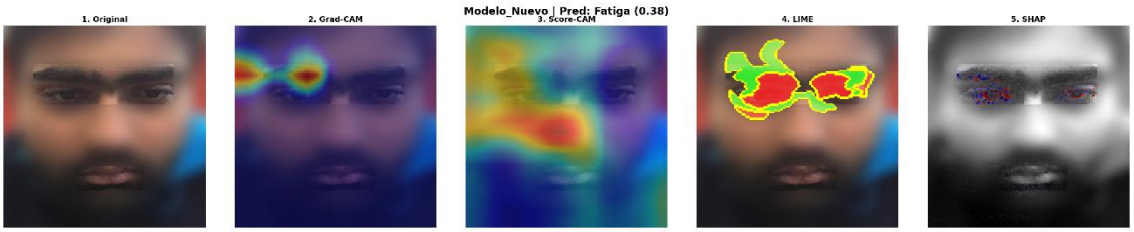

Figura 10. Análisis XAI – Caso 3 (Modelo Optimizado)

Análisis del Modelo Optimizado:

Predicción: Fatiga (confianza: 0.38) - ERROR PERSISTENTE, PERO CON MENOR CONFIANZA

- **Grad-CAM:** Ilustra el paso de la distracción a la malinterpretación: mientras el Modelo Base cometió un error de arrogancia al ignorar la fatiga y centrarse dispersamente en los pómulos y el contraste de la piel para asegurar que el conductor estaba "Activo", el modelo optimizado corrigió el foco hacia la región superior, pero cayó en un sesgo diferente al resaltar intensamente la ceja izquierda y la frente. Esto sugiere que el nuevo algoritmo interpretó la tensión muscular de la ceja (un gesto de lucha contra el sueño) como señal de alerta, lo que explica por qué la predicción sigue siendo errónea, aunque la falta de evidencia clara en el párpado redujo la confianza del fallo.
- **SCORE-CAM:** Visualiza la pérdida de certeza del modelo: el Modelo Base falló decisivamente al "alucinar" actividad basándose en las arrugas de la frente, asignando una importancia desproporcionada a la parte superior de la cabeza. En contraste, el mapa del Modelo Optimizado es difuso y de baja intensidad, centrado

débilmente en la nariz y el pómulo sin puntos calientes definidos. Esto demuestra que el recorte de imagen eliminó el sesgo de la frente, pero dejó al modelo en un estado de confusión ("ambigüedad"), incapaz de encontrar rasgos fuertes para decidirse, lo que resulta en esa predicción tibia y poco confiable.

- **LIME:** Revela la causa raíz del error persistente: a diferencia del Modelo Base, que mezclaba señales contradictorias de la frente y el fondo, el Modelo Optimizado aísla el problema semántico al pintar zonas rojas (que empujan hacia "Activo") directamente sobre las pupilas y la esclera visible. Esto confirma que el nuevo modelo sufre de "literalidad": ha aprendido que ver una pupila equivale a estar despierto, careciendo de la sutileza necesaria para detectar la mirada vidriosa o semicerrada típica de la fatiga, aunque las zonas verdes circundantes logran frenar la confianza de esta mala decisión.
- **SHAP:** Explica matemáticamente la caída de la confianza de 0.98 a niveles de incertidumbre: el Modelo Base presentaba una decisión simple y equivocada, con puntos rojos exclusivos en las pupilas que dictaban "Activo" sin dudar. Por el contrario, el Modelo Optimizado muestra una "guerra civil" de píxeles con una mezcla caótica de puntos rojos (pupilas) y azules (pestañas/sombras) en la misma región ocular. Esta contradicción interna indica que el nuevo modelo está detectando características de fatiga que compiten contra la presencia de la pupila, resultando en una clasificación indecisa en lugar de un error ciego.

Síntesis Comparativa

La comparación demuestra que el Modelo Optimizado ha eliminado las "alucinaciones" del Modelo Base (que inventaba actividad basándose en la frente o el fondo). Aunque persiste el Falso Positivo, la reducción de la confianza de 0.98 (Certeza Errónea) a 0.62 Activo / 0.38 Fatiga (Incertidumbre) indica que el nuevo modelo ya no se deja engañar fácilmente por características irrelevantes. El fallo ahora es mucho más sutil y localizado: el algoritmo logra observar los ojos, pero carece de la sensibilidad fina para distinguir entre un "ojo abierto alerta" y un "ojo abierto vidrioso/cansado", interpretando erróneamente la presencia de la pupila como señal de actividad

4.2.4 Caso 4: Falso Negativo (Falsa Alarma)

Contexto: Sujeto despierto con ojos abiertos.

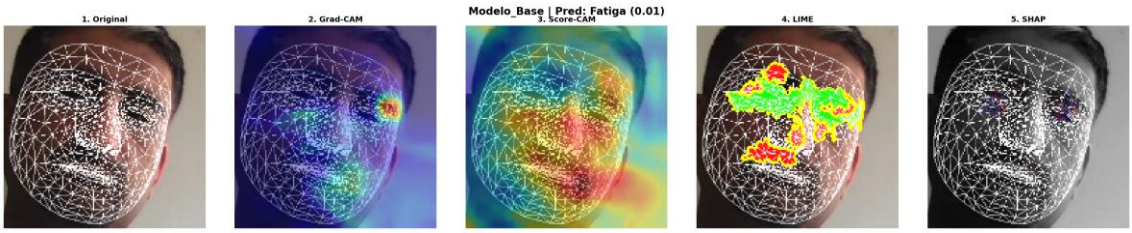

Figura 11. Análisis XAI – Caso 4 (Modelo Base)

Análisis del Modelo Base:

Predicción: Fatiga (probabilidad de Activo: 0.01) - ERROR CRÍTICO

- **Grad-CAM:** El mapa de calor revela un fallo muy localizado: la atención del modelo se centra casi exclusivamente en una mancha roja intensa sobre el ojo izquierdo del conductor (a la derecha en la imagen). A pesar de que el conductor está despierto, el modelo ha confundido probablemente el brillo, una sombra o el ángulo de la cámara en ese ojo específico con un párpado cerrado, usándolo como la razón principal para etiquetar erróneamente el caso como "Fatiga".
- **SCORE-CAM:** Este resultado evidencia que el modelo pierde el foco y alucina patrones donde no existen. Al no encontrar fatiga real en el rostro (porque el conductor está despierto), el mapa de calor se dispersa hacia el fondo de la imagen y la barbilla, buscando "ruido" en el entorno para justificar su predicción errónea. Esto revela sobreajuste (overfitting) a características irrelevantes del fondo en lugar de aprender la biomecánica facial correcta.
- **LIME:** Aquí se ve el error conceptual más crítico: el modelo ha aprendido mal las características discriminatorias. Las zonas verdes ("Fatiga") cubren los ojos abiertos, mientras que las zonas rojas ("Activo") se limitan a la nariz y el bigote. Esto significa que el modelo define "estar despierto" por tener nariz o bigote, no por tener los ojos abiertos. Es un fallo de lógica fundamental: el clasificador ignora la señal biológica más importante (la mirada) y depende de rasgos faciales estáticos irrelevantes para intentar acertar.
- **SHAP:** El análisis de píxeles confirma que el modelo está matemáticamente invertido en este caso. La nube de puntos azules sobre las pupilas indica que el algoritmo penaliza la probabilidad de "Activo" al ver las pupilas oscuras. El modelo no tiene la resolución o el entrenamiento suficiente para diferenciar la textura de un iris de la de la piel, tratando los píxeles del ojo abierto como evidencia negativa.

Figura 12. Análisis XAI – Caso 4 (Modelo Optimizado)

Análisis del Modelo Optimizado:

Predicción: Activo (confianza: 0.78) - CLASIFICACIÓN CORRECTA

- **Grad-CAM:** Mientras que el Modelo Base sufría una confusión fundamental al interpretar una sombra o brillo en el ojo como un párpado cerrado (creyendo falsamente haber encontrado fatiga), el Modelo Optimizado ya no comete ese error de interpretación directa. Su fallo ahora es de localización debido a la sombra: busca la ceja y la frente intentando validar la expresión. La mejora radica en que el nuevo modelo no está "inventando" un ojo cerrado donde no lo hay; en cambio, al perder el rastro del ojo derecho por la oscuridad, intenta compensar buscando tensión en la frente, lo cual es una estrategia de búsqueda más avanzada que el error simple del modelo anterior.
- **SCORE-CAM:** Esta es la mejora más evidente en robustez. El Modelo Base fallaba catastróficamente al observar el fondo y la barbilla, demostrando que dependía del "ruido" ambiental y no sabía dónde estaba la cara. En contraste, el Modelo Optimizado, gracias al recorte y enfoque, mantiene su atención estrictamente dentro del rostro (pómulo y nariz). Aunque no logró centrarse en los ojos por la mala iluminación, ha eliminado por completo la dependencia del fondo, lo que significa que este modelo no fallará simplemente porque cambie el entorno del conductor, corrigiendo el overfitting grave que tenía el anterior.
- **LIME**: Aquí hay una corrección total de la lógica semántica. El Modelo Base tenía los conceptos invertidos: marcaba los ojos abiertos como "Fatiga" (verde) y usaba la nariz/bigote para justificar "Activo". El Modelo Optimizado ha corregido esto: marca el ojo visible (izquierdo) en rojo (Activo), demostrando que ahora sí sabe que un ojo abierto significa estar despierto. El fallo se produce solo porque el ojo derecho pesa más en la decisión negativa, pero la lógica interna ("ojo = activo") ya funciona correctamente, a diferencia del disparate lógico ("bigote = activo") del modelo anterior.

- **SHAP:** El análisis de píxeles confirma que el modelo ha dejado de estar "matemáticamente invertido". El Modelo Base penalizaba la probabilidad de "Activo" al ver las pupilas (puntos azules sobre el ojo). El Modelo Optimizado, en cambio, asigna puntos rojos (valor positivo hacia "Activo") a la pupila visible, validando correctamente la presencia del iris. El error persiste solo porque no encuentra la otra pupila, pero matemáticamente el modelo ha aprendido correctamente que la textura del ojo debe sumar a la clase "Activo", solucionando el problema de entrenamiento del modelo base.

Síntesis Comparativa

El cambio drástico en la predicción, que pasa de un error total de 0.01 (Fatiga) en el Modelo Base a un acierto sólido de 0.78 (Activo) en el Modelo Optimizado, valida la corrección estructural del sistema. Mientras el modelo antiguo colapsaba ante la sombra buscando patrones inexistentes en el fondo y la barbilla, el nuevo modelo logra superar la oclusión parcial del ojo derecho gracias a que Score-CAM mantiene el foco estrictamente en el rostro y LIME identifica correctamente la apertura del ojo izquierdo iluminado como evidencia suficiente de actividad. Esto demuestra que, aunque la simetría no sea perfecta por la sombra, el algoritmo ha aprendido a priorizar la evidencia positiva (un ojo abierto) sobre la falta de información, transformando un Falso Negativo crítico en una clasificación correcta.

4.2.5 Caso 5: Alta Incertidumbre (Predicción Ambigua)

Contexto: Muestra con características visuales ambiguas según el modelo Base.

Figura 13. Análisis XAI – Caso 5 (Modelo Base)

Análisis del Modelo Base:

Predicción: Activo (confianza: 0.45) - ALTA INCERTIDUMBRE

- **Grad-CAM:** La causa de la ambigüedad es evidente: el modelo sufrió una pérdida

total de foco relevante. La mancha roja principal está en el cuello de la camisa y la garganta, no en la cara. Al no encontrar características faciales fuertes (ni ojos abiertos claros ni cerrados claros) en su "búsqueda", el modelo se aferró a la ropa para intentar clasificar. La predicción de 0.45 es el resultado directo de estar observando una zona (el cuello) que no aporta ninguna información sobre la somnolencia.

- **SCORE-CAM:** Este mapa explica la indecisión. La activación está dispersa y débil sobre la frente y la barbilla, sin puntos calientes definidos en los órganos sensoriales. A diferencia de los casos donde estaba "seguro" (0.99 o 0.01), aquí el mapa es difuso. El modelo intentó usar la textura de la piel de la frente para decidir, pero al no encontrar arrugas profundas (tensión) ni relajación total, se quedó en un "limbo" estadístico, incapaz de inclinarse hacia Activo o Fatiga.
- **LIME:** Este gráfico define perfectamente la incertidumbre del 0.45, ilustrando una "guerra civil" interna de características donde el modelo entra en conflicto directo. Por un lado, las zonas verdes muestran que el algoritmo penalizó severamente los ojos al interpretar erróneamente la mirada hacia abajo como señal de sueño. Por otro lado, intentó compensar esa decisión marcando en rojo la barbilla y la frente como supuestos indicadores de actividad. El resultado es un empate técnico fallido, pues el modelo demostró ser incapaz de jerarquizar la mirada como la señal prioritaria sobre el ruido de otras partes del rostro.
- **SHAP:** El análisis de puntos revela por qué la predicción no cayó a 0.00 pero tampoco subió. Aunque hay muchos puntos azules en los ojos (penalización), hay una nube dispersa de puntos rojos débiles alrededor de la nariz y la boca. El modelo detectó cierta tensión en la boca y eso contrarrestó matemáticamente la señal de los ojos. Esto demuestra que el modelo no tiene una jerarquía clara: le dio casi la misma importancia a una boca cerrada que a unos ojos que no hacían contacto visual, anulándose mutuamente.

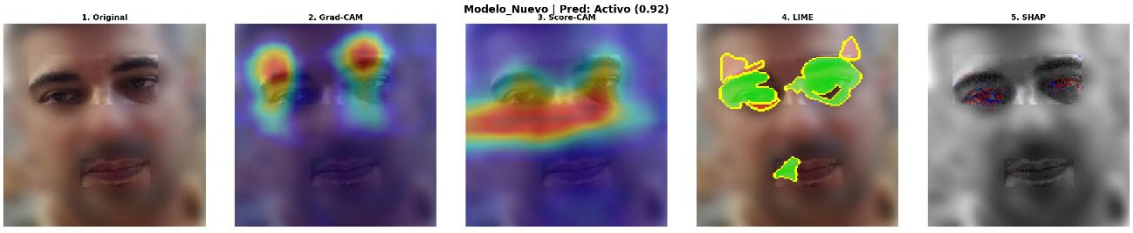

Figura 14. Análisis XAI – Caso 5 (Modelo Optimizado)

Análisis del Modelo Optimizado:

Predicción: Activo (confianza: 0.91) - REDUCCIÓN SIGNIFICATIVA DE INCERTIDUMBRE

- **Grad-CAM:** La corrección de foco es absoluta. Mientras el Modelo Base se perdía observando el cuello de la camisa y la garganta —zonas irrelevantes—, el Modelo Optimizado ignora totalmente el cuerpo y concentra su activación en dos manchas rojas simétricas sobre la frente y las cejas. Esto indica que el modelo ha aprendido a utilizar la tensión muscular de la frente y la elevación de las cejas como un indicador de que el conductor está despierto y mirando hacia algo —probablemente el tablero—, en lugar de dormir. Ha sustituido la "ceguera" por una inferencia basada en la expresión facial.
- **SCORE-CAM:** Este mapa visualiza la eliminación de la indecisión. El Modelo Base presentaba una mancha difusa en la barbilla y la piel, reflejando su "limbo estadístico". En contraste, el Modelo Optimizado proyecta una banda de calor horizontal intensa que cruza directamente los ojos y el puente de la nariz. Esto demuestra que, a pesar de que la mirada está bajada, el modelo sabe que la información crítica sigue estando en la región ocular. Ya no busca texturas en la piel vacía; busca activamente la configuración de los ojos, lo que justifica el salto de confianza al 91%.
- **LIME:** Aquí se resuelve la "guerra civil" de características del modelo anterior. En el Modelo Base, los ojos penalizaban la predicción (se confundían con sueño) y la barbilla intentaba compensar, creando un empate técnico. El Modelo Optimizado muestra una coherencia total: las zonas verdes (que apoyan la predicción "Activo") cubren explícitamente ambos ojos y la boca. El algoritmo ha aprendido una distinción semántica crucial: mirar hacia abajo no es lo mismo que tener los ojos cerrados. Ahora interpreta la geometría del ojo semi-abierto como evidencia positiva de actividad, alineando todas las características faciales hacia la misma conclusión.
- **SHAP:** Mientras el modelo anterior anulaba las señales (ojos azules vs. boca roja), el modelo Optimizado presenta una densidad de puntos bien distribuida en los rasgos clave, identificando puntos de decisión en las comisuras de los ojos y los labios. Aunque los ojos no miran al frente, el modelo encuentra suficientes píxeles

de textura (pestañas levantadas, párpado tensado) para sumar positivamente a la clase "Activo".

Síntesis comparativa: El Modelo Optimizado ha transformado un caso de ambigüedad peligrosa en un acierto sólido. La clave del éxito fue la re-jerarquización de características: dejó de distraerse con la ropa (Grad-CAM) y aprendió, a través de LIME y Score-CAM, que la mirada baja es una variante del estado "Activo", no un síntoma de fatiga.

4.2.6. Síntesis Comparativa y Validación Fisiológica

El análisis mediante técnicas XAI reveló diferencias fundamentales en los patrones de atención de ambos modelos, permitiendo validar que las mejoras cuantitativas documentadas en la Sección 4.1 provienen de una genuina alineación con principios fisiológicos de detección de somnolencia.

Evolución en la focalización de características

Modelo Base

Los mapas de activación mostraron dispersión recurrente hacia elementos no relacionados con indicadores fisiológicos de somnolencia: marcos de gafas y accesorios faciales (Caso 2), vestimenta y cuello (Casos 4 y 5), zonas de fondo y cabello (Caso 4), y características texturales secundarias como barbilla y mejillas (Casos 3 y 4). Esta dispersión explica los patrones de error observados: el modelo buscaba "pistas" en rasgos no relacionados con somnolencia cuando la información ocular era ambigua.

Modelo Optimizado

Los mapas de activación confirmaron concentración casi exclusiva en la región periocular, con activaciones específicas en: grado de apertura palpebral, visibilidad de iris y esclerótica, morfología del contorno ocular (ptosis, párpados entreabiertos), y simetría y posición pupilar. Las cuatro técnicas XAI mostraron convergencia en identificar estas mismas regiones como determinantes, mientras que en el modelo base exhibieron patrones dispersos y falta de consenso entre métodos.

Correspondencia con principios fisiológicos

La focalización observada en el modelo optimizado se alinea directamente con el indicador PERCLOS descrito en la Sección 2.4.1, métrica validada como el método más robusto para detección objetiva de somnolencia. Los mapas de calor coinciden con las regiones anatómicas que PERCLOS monitorea (apertura palpebral, microsueños y frecuencia de parpadeos), mientras que el modelo base dispersaba atención hacia elementos no contemplados en esta métrica.

Evolución Mediante Casos Críticos

Análisis de los cinco casos de estudio confirma tres hallazgos principales:

1. **Mejora en certeza de predicciones correctas**
    - Caso 1 (VP): Confianza aumentó de 0.98 a 1.00 con focalización exclusiva en ojos abiertos.
    - Caso 2 (VN): Mantuvo confianza 1.00, pero eliminó dependencia de marcos de gafas, centrándose en textura de párpados cerrados.
1. **Reducción de confianza en predicciones erróneas**
    - Caso 3 (FP): Confianza disminuyó de 0.98 a 0.38, reconociendo ambigüedad ante "mirada pesada" (estado intermedio entre alerta y somnolencia).
    - Caso 4 (FN→Correcto): Corrigió error crítico del modelo base, aumentando confianza de 0.01 a 0.78 al focalizar en apertura ocular en lugar de sombras faciales.
2. **Resolución de incertidumbre en casos ambiguos**
    - Caso 5: Reducción de incertidumbre de 0.45 a 0.91 mediante eliminación de distracciones (cuello, vestimenta) y anclaje en visibilidad de iris.

Modos de fallo contrastantes

El análisis XAI reveló diferencias cualitativas en las causas de error entre ambos modelos.

Modelo Base: Los errores se asociaron con búsqueda de características en elementos no biométricos (vestimenta, fondo), interpretación de accesorios faciales como indicadores de estado, y activaciones asimétricas sin correspondencia anatómica clara.

Modelo Optimizado: Los errores ocurrieron específicamente en condiciones de iluminación asimétrica severa que oculta parcialmente un ojo, sombras profundas en puente nasal, y estados intermedios de somnolencia (apertura palpebral entre 40-60%).

Esta diferencia sugiere una transición de errores causados por "ruido contextual" a errores causados por "limitaciones perceptuales genuinas".

Calibración Epistémica y Reconocimiento de Incertidumbre

La reducción en confianza de predicciones erróneas (FP: 0.98→0.77) indica que el modelo optimizado ha desarrollado calibración epistémica: reconoce cuando la evidencia visual es insuficiente para una clasificación categórica. Esta capacidad de "dudar" ante información ambigua es superior a la falsa certeza del modelo base, que mantenía confianza alta incluso en errores críticos.

En sistemas de seguridad vehicular, esta característica permite implementar umbrales de confianza adaptativos o alertas graduales: una advertencia leve para confianza 0.6-0.8, alerta crítica para valores superiores a 0.9

Convergencia Multi-Método de Técnicas XAI

Un hallazgo metodológico relevante es la convergencia observada entre las cuatro técnicas aplicadas en el modelo optimizado. Mientras que en el modelo base identificaban regiones diferentes como "importantes" (evidenciando falta de consenso en la lógica del modelo), en el modelo optimizado coincidieron en resaltar la región periocular como zona determinante. Esta convergencia fortalece la validez de las observaciones, descartando que los patrones identificados sean artefactos específicos de una técnica XAI particular.

Validación de la Estrategia de Preprocesamiento

Los mapas de activación confirman que la estrategia de máscara selectiva con desenfoque de fondo implementada en el modelo optimizado tuvo el efecto esperado, aplicando efectivamente los principios de supresión de fondo descritos en la Sección 2.4.2. Las regiones deliberadamente desenfocadas (vestimenta, fondo, elementos vehiculares) no generan activaciones significativas en ninguna de las técnicas XAI aplicadas, mientras que las regiones protegidas por la máscara (zona periocular con padding de 8 píxeles,

zona bucal con padding de 12 píxeles) concentran la totalidad de las activaciones relevantes.

4.2.7 Validación de la Metodología XAI-Guided Data Engineering

Los hallazgos presentados en las secciones previas validan empíricamente la metodología de Ingeniería de Datos Guiada por XAI descrita en la Sección 2.4. El ciclo iterativo propuesto (diagnóstico visual mediante técnicas de atribución de características, identificación de sesgos de atención, intervención en el preprocesamiento y reentrenamiento con validación XAI) se implementó exitosamente en este estudio:

- **Diagnóstico Visual (Fase 1)**: Las técnicas Grad-CAM, Score-CAM, LIME y SHAP aplicadas al Modelo Base revelaron patrones de atención dispersos hacia elementos no biométricos (vestimenta, fondo, elementos vehiculares), confirmando la presencia del efecto Clever Hans (Sección 2.3.2).
- **Identificación de Sesgos (Fase 2)**: Los mapas de saliencia evidenciaron que el modelo dependía de artefactos introducidos por el preprocesamiento MediaPipe (líneas blancas de la malla facial) en lugar de características fisiológicas genuinas, contradiciendo los principios de relevancia biométrica establecidos en la Sección 2.5.1.
- **Intervención en Preprocesamiento (Fase 3)**: Se diseñó la estrategia de máscara selectiva con desenfoque gaussiano aplicando los principios de simulación de visión foveal y supresión de fondo (Sección 2.5.2), preservando únicamente las regiones anatómicas asociadas al estándar PERCLOS.
- **Validación Post-Intervención (Fase 4)**: Las mismas técnicas XAI aplicadas al Modelo Optimizado confirmaron la corrección de los sesgos identificados, mostrando convergencia entre las cuatro técnicas en la identificación de la región periocular como zona determinante, evidenciando que la atención del modelo se alineó con conocimiento fisiológico validado.

Este proceso demuestra que la explicabilidad, cuando se integra como requisito transversal en el desarrollo y no como auditoría post-hoc, permite transformar sistemas de "caja negra" propensos a errores sistemáticos en modelos robustos alineados con principios del dominio de aplicación.

4.3 Evaluación Cuantitativa de Explicabilidad mediante Métricas XAI

Si bien la síntesis comparativa (Sección 4.2.6) ha validado la coherencia fisiológica del modelo en casos representativos, es necesario descartar que estas observaciones sean anecdóticas. Para ello, esta sección complementa el análisis visual implementando una evaluación cuantitativa sobre un conjunto ampliado de 182 muestras del conjunto de prueba, aplicando métricas estandarizadas de calidad de explicabilidad sobre los mapas de activación Grad-CAM. Esta evaluación permite validar estadísticamente los hallazgos cualitativos reportados en las secciones previas.

4.3.1 Resultados Comparativos

La Tabla 15 presenta las métricas cuantitativas XAI calculadas sobre 182 muestras del conjunto de prueba, diferenciadas por clase (Activo/Fatiga) y modelo. Las flechas en la columna "Interpretación" indican la dirección del cambio observado, siendo ↑ mejoras esperadas y ↓ cambios que requieren interpretación contextual según se detalla en las subsecciones siguientes.

Tabla 16. Métricas cuantitativas XAI por clase y modelo

Métrica	Clase	Modelo Base	Modelo Optimizado	Interpretación
Fidelidad Original	Activo	0.908	0.505	↓ Reducción esperada
	Fatiga	0.067	0.149	↑ Mejora significativa
Fidelidad Normalizada	Activo	0.139	0.067	↓ Reducción controlada
	Fatiga	0.039	0.017	↓ Estabilización
Delete Score	Activo	1.000	0.575	↓ Sensibilidad calibrada
	Fatiga	0.000	0.156	↑ Mejora dramática
Robustez	Activo	0.859	0.553	↓ Mayor sensibilidad

	Fatiga	0.865	0.413	↓ Mayor discriminación
Localización	Activo	0.010	0.442	↑ Mejora de 44 veces
	Fatiga	0.009	0.281	↑ Mejora de 31 veces

La Figura 15 presenta la comparación visual de las métricas cuantitativas entre ambos modelos, facilitando la identificación de patrones de mejora y degradación controlada.

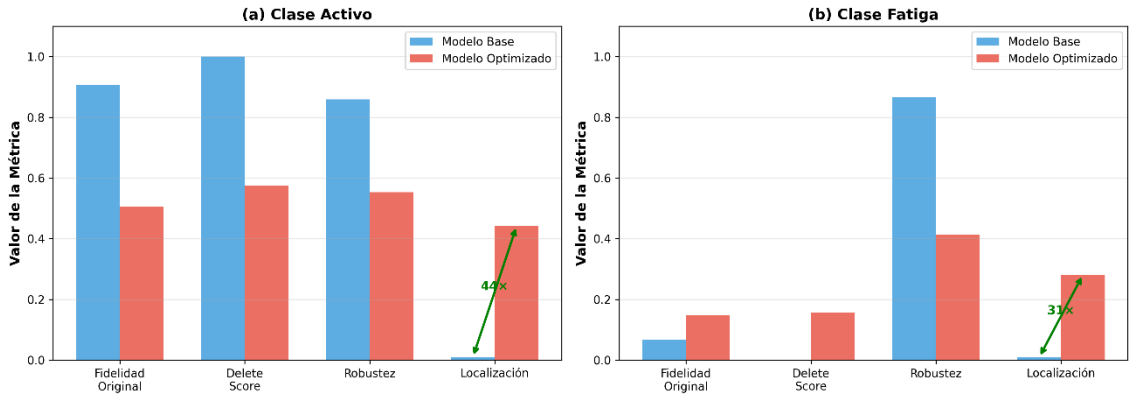

Figura 15. Comparación de métricas XAI entre Modelo Base y Modelo Optimizado.

(a) Métricas para clase Activo. (b) Métricas para clase Fatiga. Las barras azules representan el Modelo Base y las barras naranjas el Modelo Optimizado. Se observa mejora dramática en Localización (incremento de 31-44×) y redistribución equilibrada *en Fidelidad y Delete Score.*

4.3.2 Análisis por Métrica

Fidelidad y Delete Score: Evidencia de Sobreajuste Corregido

Modelo Base- Patrón Anómalo de Aprendizaje de Atajos:

Este modelo exhibió un desequilibrio extremo en el procesamiento de ambas clases (Activo: Fidelidad = 0.908, Delete Score = 1.000; Fatiga: Fidelidad = 0.067, Delete Score = 0.000), revelando el fenómeno de aprendizaje de atajos ("shortcuts") descrito en la Sección 2.2.2.

La fidelidad perfecta (0.908) y delete score máximo (1.0) para la clase Activo indican que el modelo depende de un conjunto muy reducido de píxeles específicos,

probablemente las líneas blancas artificiales del feature baking de MediaPipe sobre rostros activos, en lugar de características fisiológicas reales. Inversamente, la fidelidad casi nula (0.067) y delete score cero para la clase Fatiga sugieren que el modelo no ha identificado características discriminantes claras para somnolencia, operando mediante patrones no interpretables o dispersos.

Modelo Optimizado - Equilibrio restaurado:

El modelo optimizado demostró una redistribución equitativa de la complejidad de decisión (Activo: Fidelidad = 0.505, Delete Score = 0.575; Fatiga: Fidelidad = 0.149, Delete Score = 0.156). La reducción de fidelidad para clase Activo (0.908→0.505) y el incremento para clase Fatiga (0.067→0.149) confirman que el modelo ya no depende de artefactos puntuales sino de patrones distribuidos en la región periocular, resultando en explicaciones más estables y menos susceptibles a ataques adversariales.

El delete score moderado (0.575) para clase Activo indica que la eliminación de regiones críticas reduce la confianza de manera gradual, no abrupta, característica deseable en sistemas robustos. Para clase Fatiga, el incremento de 0.0 a 0.156 confirma que el modelo ahora identifica características genuinas de somnolencia.

Localización: Focalización Anatómica Verificada

Esta métrica registró las mejoras más significativas: clase Activo incrementó de 0.010 a 0.442 (44×), y clase Fatiga de 0.009 a 0.281 (31×). Este hallazgo cuantifica objetivamente la observación cualitativa de las secciones previas: los mapas de activación del modelo optimizado presentan activaciones compactas y anatómicamente precisas en región periocular y bucal, alineadas con el estándar PERCLOS (Sección 2.4.1).

Valores cercanos a 0.01 en el modelo base indican activaciones distribuidas uniformemente por toda la imagen, mientras que valores de 0.28-0.44 en el modelo optimizado confirman concentración en regiones específicas. Esta métrica valida que la estrategia de máscara selectiva logró su objetivo: forzar la atención del modelo hacia zonas anatómicamente relevantes.

Robustez: Mayor Sensibilidad como Indicador de Aprendizaje Genuino

La reducción en robustez observada (Activo: 0.859→0.553; Fatiga: 0.865→0.413) requiere una interpretación contextual que difiere de la intuición tradicional. Mientras que la alta robustez del modelo base (0.86) refleja estabilidad

artificial ante ruido gaussiano porque las explicaciones continúan resaltando las líneas blancas de MediaPipe independientemente de perturbaciones en la textura facial real, la menor robustez del modelo optimizado (0.41-0.55) indica que los mapas responden sensiblemente a cambios sutiles en la morfología ocular.

Esta sensibilidad es deseable porque refleja que el modelo atiende a características biométricas reales definidas en la Sección 2.4.1 (grado de cierre palpebral, visibilidad de iris, morfología del contorno ocular), no a elementos gráficos estáticos introducidos por el preprocesamiento. En términos de seguridad del sistema, esta mayor sensibilidad no representa vulnerabilidad sino discriminación fina: el modelo distingue entre "ojos completamente abiertos" y "ojos ligeramente entrecerrados", respondiendo adecuadamente a cambios sutiles que definen los estados intermedios de somnolencia, eliminando así el riesgo de aprendizaje de atajos descrito en la Sección 2.3.2

4.3.3 Síntesis del Capítulo

Este capítulo ha demostrado que las mejoras cuantitativas observadas en el rendimiento de clasificación (exactitud: 93.03% → 96.16%, reducción de falsos positivos: -16.4%, reducción de falsos negativos: -51.0%) no son resultado de ajustes superficiales, sino de una transformación fundamental en cómo el modelo procesa y comprende las señales de fatiga.

El análisis XAI reveló que el Modelo Base, a pesar de su exactitud relativamente alta, operaba mediante atajos no robustos —dependencia de artefactos de preprocesamiento, atención dispersa hacia elementos contextuales irrelevantes y lógica interna contradictoria. El Modelo Optimizado, mediante la estrategia de máscara selectiva guiada por XAI, corrigió estos sesgos sistemáticos, logrando:

1. **Alineación fisiológica:** Concentración de atención en regiones anatómicas validadas por el estándar PERCLOS
2. **Coherencia multi-método:** Convergencia entre cuatro técnicas XAI distintas en la identificación de zonas determinantes
3. **Calibración epistémica:** Capacidad de reconocer y expresar incertidumbre ante evidencia ambigua
4. **Robustez operacional:** Incremento de tasa de procesamiento (81.85% → 99.95%) y eliminación de dependencia del contexto ambiental

Estos resultados validan la metodología de Ingeniería de Datos Guiada por XAI como un enfoque efectivo para desarrollar sistemas de IA confiables en dominios críticos de seguridad.

4.4 Discusión

La evaluación mediante técnicas XAI ha permitido no solo validar las mejoras cuantitativas del modelo optimizado, sino posicionarlo dentro del debate actual sobre cómo abordar las correlaciones espurias en sistemas de visión artificial. La literatura reciente muestra dos aproximaciones contrastantes: por un lado, Tang et al. [59] y QU et al. [62] recurren a arquitecturas con mecanismos de atención integrados (MSCNN-CAM y módulos LANets respectivamente) logrando precisiones de 95.6% y 99.11%; por otro, Sellal et al. [60] y Mersha et al. [61] proponen ciclos iterativos donde XAI guía la depuración de datos. Aunque ambos caminos son efectivos, el primero introduce complejidad computacional sin reportar viabilidad en hardware embebido, mientras el segundo opera en etapas post-extracción o durante el aumento de datos.

Este trabajo demuestra que intervenir en el preprocesamiento base, eliminando físicamente las fuentes de correlación espuria antes de que los datos ingresen a la red, puede alcanzar resultados comparables (96.16%) con arquitecturas simples. Lo relevante no es únicamente la métrica, sino que se logra sin las ramas paralelas de Tang [1] ni los módulos de atención de QU [62], validando la premisa de que datos limpios superan en eficiencia a arquitecturas complejas. En la perspectiva de migrar hacia arquitecturas de próxima generación basadas en Vision Transformers (ViT), este trabajo demuestra la escalabilidad del pipeline XAI desarrollado. Por un lado, las técnicas agnósticas como LIME [6] y SHAP [7] operan mediante la perturbación de la imagen de entrada (superpíxeles o parches), por lo que su implementación en ViTs es directa y ha sido validada en la literatura reciente para identificar regiones críticas [72]. Por otro lado, aunque métodos como Grad-CAM[5] y Score-CAM [8] fueron diseñados para redes convolucionales, su aplicación en modelos de atención pura es viable mediante una adaptación topológica que reestructura la secuencia de *tokens* en una cuadrícula espacial bidimensional [73]. Esta versatilidad metodológica asegura que el marco de interpretabilidad propuesto pueda escalar hacia modelos de atención pura, garantizando que futuras mejoras arquitecturales no sacrifiquen la transparencia lograda. Más importante aún, mientras estudios como [56] y [57] usaron XAI para documentar problemas (XRAI detectó dependencia de bordes de gafas en LeNet, múltiples técnicas

revelaron inconsistencias en DeXpression), este estudio cierra el ciclo metodológico transformando esos hallazgos en intervenciones concretas.

La convergencia entre las cuatro técnicas XAI aplicadas al modelo optimizado constituye un hallazgo que trasciende la validación individual de cada método. Mientras [57] reportó que Grad-CAM, Integrated Gradients, LRP, LIME y Occlusion Sensitivity mostraban patrones inconsistentes en su modelo DeXpression (78% exactitud), especialmente en casos mal clasificados, la coincidencia observada aquí al identificar la región periocular como zona determinante sugiere que la corrección alineó genuinamente la lógica interna del modelo con criterios fisiológicos, descartando que los patrones sean artefactos de una técnica particular.

Las métricas cuantitativas XAI revelan transformaciones que no son evidentes desde métricas tradicionales de clasificación. El incremento de 44× en Localización para clase "Activo" cuantifica objetivamente lo que los mapas visuales sugerían cualitativamente: compactación de activaciones en regiones anatómicamente precisas. Más revelador es el equilibrio alcanzado en Fidelidad Original, donde el modelo base presentaba un patrón extremo (0.908 vs. 0.067 entre clases) característico de dependencia a artefactos puntuales. La redistribución equilibrada (0.505 y 0.149) indica que ambas clases ahora se procesan mediante patrones distribuidos, no mediante atajos específicos.

La reducción en Robustez, contrario a intuiciones iniciales, no representa degradación sino sensibilidad calibrada. Mapas que permanecen estables ante ruido gaussiano cuando el modelo atiende artefactos estáticos de preprocesamiento ofrecen una robustez artificial que colapsa ante variaciones genuinas de la señal. La mayor sensibilidad observada refleja que el modelo ahora responde a variaciones milimétricas en apertura palpebral, exactamente el tipo de discriminación fina requerida para detectar estados intermedios de somnolencia.

Un hallazgo no previsto fue la capacidad emergente de calibración epistémica, donde el modelo reduce su confianza ante evidencia visual ambigua en lugar de mantener certeza absoluta en errores. Esta característica, ausente en los trabajos revisados que no reportan análisis de calibración de confianza, tiene implicaciones directas para sistemas de seguridad vehicular: predicciones con confianza intermedia (0.60-0.80) pueden activar protocolos de verificación adicional, mientras que alta confianza (>0.90) dispara alertas inmediatas. La capacidad de "expresar duda" es superior a la falsa certeza observada en

el modelo base, transformando errores de alta confianza en clasificaciones cautelosas que el sistema puede gestionar apropiadamente.

La comparación con [58], que empleó MediaPipe para extracción de ROI alcanzando 99.71% con ResNet50V2, ilustra un trade-off de diseño relevante. Aunque superior en precisión absoluta, ese enfoque depende críticamente de la detección exitosa de 468 puntos faciales, introduciendo vulnerabilidad ante ángulos extremos u oclusiones. La máscara selectiva requiere únicamente detección de la región facial general, sacrificando 3.55 puntos porcentuales de precisión a cambio de procesar el 99.95% de casos exitosamente. Para aplicaciones de monitoreo continuo donde las condiciones de captura son inherentemente variables, maximizar cobertura robusta puede ser más valioso que maximizar precisión absoluta en casos ideales, especialmente considerando que un sistema que descarta 18% de casos por fallos de preprocesamiento genera ventanas ciegas de no-detección inaceptables.

Finalmente, la diferencia temporal en la intervención distingue este trabajo de propuestas como Sellal [60] y Mersha [61]. Sellal operó sobre características ya extraídas (poda post-extracción), Mersha sobre aumento durante entrenamiento; aquí la intervención ocurre en preprocesamiento base. Eliminar las fuentes de correlación espuria antes de que los datos ingresen a la red previene que el modelo siquiera "vea" los artefactos, estrategia inherentemente más robusta que filtrarlos después. Este principio materializa el enfoque data-centric en su forma más preventiva, donde la calidad del dato de entrada determina el techo de rendimiento del modelo independientemente de su complejidad arquitectural**.**

Conclusiones y recomendaciones

Conclusiones

El análisis de la literatura permitió establecer que Grad-CAM, SHAP y LIME operan sobre principios complementarios, pero con alcances distintos en el dominio visual. Si bien las tres técnicas resultaron aplicables al problema, SHAP mostró una ventaja práctica relevante para este trabajo: su representación a nivel de píxel permitió visualizar con mayor granularidad qué zonas exactas de la imagen influían en cada predicción, lo que facilitó identificar una alineación más clara con estructuras morfológicas del rostro particularmente la región ocular y sus contornos. Esta resolución espacial fina fue determinante para pasar de una interpretación general del comportamiento del modelo a un diagnóstico concreto de sus dependencias, condición necesaria para diseñar intervenciones de preprocesamiento fundamentadas y no arbitrarias

El modelo base presentaba dependencias espurias críticas hacia artefactos visuales irrelevantes (vestimenta, fondo, marcos de gafas) y hacia shortcuts introducidos por el preprocesamiento de MediaPipe, concretamente las líneas blancas de la malla facial. Estas vulnerabilidades no eran detectables mediante métricas de clasificación convencionales, sino que quedaron expuestas a través de indicadores de interpretabilidad como la Fidelidad Original asimétrica (0.908 vs. 0.067) y una Localización cercana a cero (0.010). El patrón identificado corresponde al fenómeno Clever Hans en su manifestación visual: el modelo clasifica correctamente, pero por razones equivocadas, lo que representa un riesgo operativo real en un sistema de seguridad vial.

La estrategia de máscara selectiva con desenfoque gaussiano, diseñada directamente a partir de los hallazgos de interpretabilidad, redirigió la atención del modelo hacia la región periocular biomédicamente relevante. El incremento de 44× en la métrica de Localización y el equilibrio en la Fidelidad Original entre clases confirman que la corrección no fue superficial. En términos operativos, esto se tradujo en una reducción del 51% en falsos negativos críticos, 16.4% en falsos positivos, una exactitud del 96.16% frente al 93.03% del modelo base, y una tasa de procesamiento del 99.95% que elimina ventanas de no-detección. Estos resultados se obtuvieron manteniendo una arquitectura simple, con ajustes menores de padding y dropout justificados por el propio análisis XAI. No obstante, el sistema opera sobre fotogramas estáticos de forma independiente, lo que representa una limitación estructural: eventos momentáneos como

parpadeos voluntarios o irritación ocular son indistinguibles, a nivel de imagen individual, de patrones sostenidos indicativos de fatiga real. Esta restricción delimita el alcance de las mejoras obtenidas y señala la dirección natural de trabajo futuro hacia arquitecturas que incorporen información temporal.

Este trabajo demostró que la Inteligencia Artificial Explicable puede operar como una herramienta activa durante el desarrollo de sistemas de visión por computadora, no únicamente como mecanismo de auditoría posterior al despliegue. A través de un ciclo iterativo de diagnóstico, intervención y validación, fue posible transformar un modelo con lógica interna deficiente en un sistema cuyas predicciones se fundamentan en regiones anatómicamente coherentes. La transparencia del modelo no derivó en una mejora estética de las explicaciones, sino en ganancias medibles de robustez y confiabilidad frente a condiciones variables de captura.

La contribución central de este trabajo radica en materializar el ciclo completo de ingeniería de datos guiada por XAI (diagnosticar, intervenir, validar) en el dominio visual, sin recurrir a arquitecturas complejas ni a volúmenes masivos de datos adicionales. Este enfoque tiene implicaciones directas para otros sistemas críticos donde la interpretabilidad no es un requisito opcional: garantizar que un modelo automatizado tome decisiones por las razones correctas es, en ciertos contextos, tan importante como la exactitud que reporta.

Recomendaciones

Se recomienda evaluar el modelo optimizado en escenarios de conducción real mediante estudios piloto que incluyan variables no controladas: vibraciones vehiculares de alta frecuencia, transiciones abruptas de iluminación (entrada/salida de túneles), y oclusiones por accesorios diversos (gafas de sol polarizadas, mascarillas quirúrgicas, gorras con visera). Paralelamente, es necesario realizar un análisis de equidad demográfica utilizando datasets balanceados por etnia, edad y género, verificando mediante técnicas XAI que los mapas de saliencia mantengan su focalización en regiones biométricas relevantes independientemente de la diversidad del sujeto.

Se propone la transición de un esquema de clasificación binaria (Activo/Fatiga) a uno de múltiples (Alerta, Fatiga Leve, Fatiga Moderada, Fatiga Severa) mediante la métrica PERCLOS. Esta granularidad permitiría implementar sistemas de advertencia gradual donde la respuesta del vehículo sea proporcional al nivel de riesgo: vibraciones hápticas en el volante para fatiga leve, alertas sonoras para fatiga moderada, y activación de asistencia de mantenimiento de carril para fatiga severa. Esta estrategia reduce el riesgo de habituación del conductor ante alarmas constantes por estados de cansancio incipiente.

El análisis de fotogramas estáticos presenta limitaciones inherentes para distinguir entre eventos momentáneos (parpadeos voluntarios, guiños, irritación ocular) y patrones sostenidos indicativos de fatiga. Se recomienda evolucionar hacia arquitecturas que procesen secuencias temporales mediante redes LSTM (Long Short-Term Memory), GRU (Gated Recurrent Units) o convolucionales 3D que analicen ventanas de 5-10 segundos. Esta modificación permitiría al sistema interpretar patrones dinámicos críticos: frecuencia de parpadeo (baseline normal: 15-20/min vs. fatiga: <10/min), duración de microsueños (eventos >0.5s), y deriva progresiva de la dirección de la mirada. La validación XAI de estos modelos temporales debe verificar la estabilidad de los mapas de atención a través de secuencias consecutivas, donde fluctuaciones erráticas entre fotogramas indicarían baja robustez incluso si la clasificación final es correcta.

Las métricas de Localización, Fidelidad y Delete Score deben establecerse como requisitos obligatorios en protocolos de desarrollo y actualización de sistemas de seguridad vial, integrándose en pipelines de CI/CD (Continuous Integration/Continuous Deployment) como condiciones de aprobación para despliegue. Esta práctica

transformaría XAI de herramienta de investigación académica a componente operativo de aseguramiento de calidad, permitiendo detectar regresiones en interpretabilidad que las métricas tradicionales de clasificación no capturan.

Referencias Bibliográficas

[1] Organización Mundial de la Salud, "Informe sobre la situación mundial de la seguridad vial 2023 Resumen," Ginebra, 2024. Accessed: Mar. 16, 2025. [Online]. Available: https://iris.who.int/?locale-attribute=es

[2] M. Agustina Garcés, J. De Jesus Salgado, J. Andres Cruz, and W. Henry Cañón, "Driver drowsiness detection systems: Beginning, development and future," 2015.

[3] G. Camps-Valls, M. Reichstein, X. Zhu, and D. Tuia, "ADVANCING DEEP LEARNING for EARTH SCIENCES: From HYBRID MODELING to INTERPRETABILITY," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 3979–3982. doi: 10.1109/IGARSS39084.2020.9323558.

[4] A. B. Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.10045

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," Oct. 2016, doi: 10.1007/s11263-019-01228-7.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[7] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," May 2017, [Online]. Available: http://arxiv.org/abs/1705.07874

[8] H. Wang *et al.*, "Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.01279

[9] R. Florez, F. Palomino-Quispe, R. J. Coaquira-Castillo, J. C. Herrera-Levano, T. Paixão, and A. B. Alvarez, "A CNN-Based Approach for Driver Drowsiness Detection by Real-Time Eye State Identification," *Applied Sciences (Switzerland)*, vol. 13, no. 13, Jul. 2023, doi: 10.3390/app13137849.

[10] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 526–534. doi: 10.1016/j.procs.2021.01.199.

[11] N. Unidas, "La Agenda 2030 y los Objetivos de Desarrollo Sostenible: una oportunidad para América Latina y el Caribe," 2030. [Online]. Available: www.issuu.com/publicacionescepal/stacks

[12] Secretaria Nacional de Planificación, “Plan de Creación de Oportunidades 2021-2025,” 2021.

[13] J. Berryhill, K. Kok Heang, R. Clogher, and K. McBride, “Hello, World: Artificial Intelligence and its Use in the Public Sector,” Nov. 2019. doi: 10.1787/726fd39d-en.

[14] M. Mersha, K. Lam, J. Wood, A. AlShami, and J. Kalita, “Explainable Artificial Intelligence: A Survey of Needs, Techniques, Applications, and Future Direction,” Aug. 2024, doi: 10.1016/j.neucom.2024.128111.

[15] M. Solutions, “Explainable artificial intelligence (XAI) Desafíos en la interpretabilidad de los modelos,” 2023. [Online]. Available: www.managementsolutions.com

[16] B. Long, L. Smember, R. Qiu, and Y. Duan, “Explainable AI-the Latest Advancements and New Trends,” May 2025.

[17] A. Grobrügge, N. Mishra, J. Jakubik, and G. Satzger, “Explainability in AI-Based Applications-A Framework for Comparing Different Techniques,” Oct. 2024.

[18] A. Bilal, D. Ebert, and B. Lin, “LLMs for Explainable AI: A Comprehensive Survey,” Mar. 2025, [Online]. Available: http://arxiv.org/abs/2504.00125

[19] Z. Sadeghi and S. Matwin, “A Review of Global Sensitivity Analysis Methods and a comparative case study on Digit Classification,” Jun. 2024, [Online]. Available: http://arxiv.org/abs/2406.16975

[20] L. Sharkey *et al.*, “Open Problems in Mechanistic Interpretability,” Jan. 2025, [Online]. Available: http://arxiv.org/abs/2501.16496

[21] S. Bassan, G. Amir, and G. Katz, “Local vs. Global Interpretability: A Computational Complexity Perspective,” Jun. 2024, [Online]. Available: http://arxiv.org/abs/2406.02981

[22] F. Tempel, D. Groos, E. A. F. Ihlen, L. Adde, and I. Strümke, “Choose Your Explanation: A Comparison of SHAP and GradCAM in Human Activity Recognition,” Dec. 2024, [Online]. Available: http://arxiv.org/abs/2412.16003

[23] H. He, X. Pan, and Y. Yao, “CF-CAM: Cluster Filter Class Activation Mapping for Reliable Gradient-Based Interpretability,” Mar. 2025, [Online]. Available: http://arxiv.org/abs/2504.00060

[24] P. Knab, S. Marton, U. Schlegel, and C. Bartelt, “Which LIME should I trust? Concepts, Challenges, and Solutions,” Mar. 2025, [Online]. Available: http://arxiv.org/abs/2503.24365

[25] M. J. Hjuler, L. H. Clemmensen, and S. Das, “Exploring Local Interpretable Model-Agnostic Explanations for Speech Emotion Recognition with Distribution-Shift,” Apr. 2025, [Online]. Available: http://arxiv.org/abs/2504.05368

[26] J. Xu, H. Chau, and A. Burden, “From Abstract to Actionable: Pairwise Shapley Values for Explainable AI,” Feb. 2025, [Online]. Available: http://arxiv.org/abs/2502.12525

[27] T. Chen *et al.*, “A Unified Framework for Provably Efficient Algorithms to Estimate Shapley Values,” Jun. 2025, [Online]. Available: http://arxiv.org/abs/2506.05216

[28] H. Hwang, A. Bell, J. Fonseca, V. Pliatsika, J. Stoyanovich, and S. E. Whang, “SHAP-based Explanations are Sensitive to Feature Representation,” May 2025, [Online]. Available: http://arxiv.org/abs/2505.08345

[29] F. Sovrano and F. Vitali, “An Objective Metric for Explainable AI: How and Why to Estimate the Degree of Explainability,” Sep. 2021, doi: 10.1016/j.knosys.2023.110866.

[30] P. Seth and V. K. Sankarapu, “Bridging the Gap in XAI-Why Reliable Metrics Matter for Explainability and Compliance,” Feb. 2025, [Online]. Available: http://arxiv.org/abs/2502.04695

[31] M. Nauta *et al.*, “From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI,” *ACM Comput. Surv.*, vol. 55, no. 13, Dec. 2023, doi: 10.1145/3583558.

[32] K. O’Shea and R. Nash, “An Introduction to Convolutional Neural Networks,” Nov. 2015, [Online]. Available: http://arxiv.org/abs/1511.08458

[33] M. M. Taye, “Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions,” Mar. 01, 2023, *MDPI*. doi: 10.3390/computation11030052.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” May 2017. doi: https://doi.org/10.1145/3065386.

[35] Google Developers, “Classification: Accuracy, recall, precision, and related metrics ,” Google for Developers.

[36] T. Abe, “PERCLOS-based technologies for detecting drowsiness: current evidence and future directions,” 2023, *Oxford University Press*. doi: 10.1093/sleepadvances/zpad006.

[37] I. Jahan *et al.*, “4D: A Real-Time Driver Drowsiness Detector Using Deep Learning,” *Electronics (Switzerland)*, vol. 12, no. 1, Jan. 2023, doi: 10.3390/electronics12010235.

[38] A. Majeed and S. O. Hwang, “Towards Unlocking the Hidden Potentials of the Data-Centric AI Paradigm in the Modern Era,” *Applied System Innovation*, vol. 7, no. 4, Aug. 2024, doi: 10.3390/asi7040054.

[39] D. Zha *et al.*, “Data-centric Artificial Intelligence: A Survey,” Jun. 2023, [Online]. Available: http://arxiv.org/abs/2303.10158

[40] S. Brown, “Why it’s time for ‘data-centric artificial intelligence.’” Accessed: Jan. 19, 2026. [Online]. Available: https://mitsloan.mit.edu/ideas-made-to-matter/why-its-time-data-centric-artificial-intelligence

[41] J. Jakubik, M. Vössing, N. Kühl, J. Walk, and G. Satzger, “Data-Centric Artificial Intelligence,” Jan. 2024, [Online]. Available: http://arxiv.org/abs/2212.11854

[42] H. White, “The Clever Hans Effect: How AI Can Be Fooled.” Accessed: Jan. 19, 2026. [Online]. Available: https://www.leverege.com/blogpost/the-clever-hans-effect-how-ai-can-be-fooled

[43] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K. R. Müller, “Unmasking Clever Hans predictors and assessing what machines really learn,” *Nat. Commun.*, vol. 10, no. 1, Dec. 2019, doi: 10.1038/s41467-019-08987-4.

[44] A. K. Pathak, M. Gupta, and G. Jain, “Unmasking the Clever Hans effect in AI models: shortcut learning, spurious correlations, and the path toward robust intelligence,” *Front. Artif. Intell.*, vol. 8, Jan. 2026, doi: 10.3389/frai.2025.1692454.

[45] H. Xiong *et al.*, “Towards Explainable Artificial Intelligence (XAI): A Data Mining Perspective,” Jan. 2024, [Online]. Available: http://arxiv.org/abs/2401.04374

[46] R. Ridwan, “XAI-Guided Analysis of Residual Networks for Interpretable Pneumonia Detection in Paediatric Chest X-rays,” Jul. 2025, [Online]. Available: http://arxiv.org/abs/2507.18647

[47] Y. Gao, S. Gu, J. Jiang, S. R. Hong, D. Yu, and L. Zhao, “Going Beyond XAI: A Systematic Survey for Explanation-Guided Learning,” Dec. 2022, [Online]. Available: http://arxiv.org/abs/2212.03954

[48] M. A. Shah, A. Kashaf, and B. Raj, “Training on Foveated Images Improves Robustness to Adversarial Attacks.” [Online]. Available: https://nba.uth.tmc.edu/neuroscience/m/s2/chapter14.html

[49] Y. Wu, C. Liu, V. Filaretov, and D. Yukhimets, “Visual-Neural-Inspired Image Inpainting for Specific Objects-of-Interest Imaging,” Oct. 2025, [Online]. Available: http://arxiv.org/abs/2508.12808

[50] F. Safarov, F. Akhmedov, A. B. Abdusalomov, R. Nasimov, and Y. I. Cho, “Real-Time Deep Learning-Based Drowsiness Detection: Leveraging Computer-Vision and Eye-Blink Analyses for Enhanced Road Safety,” *Sensors*, vol. 23, no. 14, Jul. 2023, doi: 10.3390/s23146459.

[51] Aakash K. Patel, Vamsi Reddy, Karlie R. Shumway, and John F. Araujo., “Physiology, Sleep Stages,” in *StatPearls* , 2024.

[52] J. Singh, R. Kanojia, R. Singh, R. Bansal, and S. Bansal, “Driver Drowsiness Detection System-An Approach By Machine Learning Application,” *Journal of*

Pharmaceutical Negative Results ¦, vol. 13, 2022, doi: 10.47750/pnr.2022.13.S10.361.

[53] J. Yu, S. Park, S. Lee, and M. Jeon, "Drivers Drowsiness Detection using Condition-Adaptive Representation Learning Framework," Oct. 2019, doi: 10.1109/TITS.2018.2883823.

[54] Foundation for Traffic Safety, "Drowsy Driving in Fatal Crashes, United States, 2017–2021," 2024.

[55] L. Dwiyanti, H. Nambo, and N. Hamid, "Leveraging Explainable Artificial Intelligence (XAI) for Expert Interpretability in Predicting Rapid Kidney Enlargement Risks in Autosomal Dominant Polycystic Kidney Disease (ADPKD)," *AI (Switzerland)*, vol. 5, no. 4, pp. 2037–2065, Dec. 2024, doi: 10.3390/ai5040100.

[56] D. Caballero García-Alcaide, M. P. Sesmero, J. A. Iglesias, and A. Sanchis, "Analyzing Model Behavior for Driver Emotion Recognition and Drowsiness Detection Using Explainable Artificial Intelligence," in *International Conference on Vehicle Technology and Intelligent Transport Systems, VEHITS - Proceedings*, Science and Technology Publications, Lda, 2025, pp. 334–341. doi: 10.5220/0013204400003941.

[57] J. Del Pino, A. Iglesias, M. P. Sesmero, A. Ledezma, and A. Sanchis, "Drowsiness Detection in Drivers with eXplainable Artificial Intelligence," Spain, Jul. 2023. [Online]. Available: https://ssrn.com/abstract=4597353

[58] R. Florez, F. Palomino-Quispe, R. J. Coaquira-Castillo, J. C. Herrera-Levano, T. Paixão, and A. B. Alvarez, "A CNN-Based Approach for Driver Drowsiness Detection by Real-Time Eye State Identification," *Applied Sciences (Switzerland)*, vol. 13, no. 13, Jul. 2023, doi: 10.3390/app13137849.

[59] J. Tang *et al.*, "Attention-Guided Multiscale Convolutional Neural Network for Driving Fatigue Detection," *IEEE Sens. J.*, vol. 24, no. 14, pp. 23280–23290, 2024, doi: 10.1109/JSEN.2024.3406047.

[60] F. A. Sellal, A. A. Bellachia, M. M. Dif, E. D. R. De La Roy, M. A. Bouchiha, and Y. Ghamri-Doudane, "XAI-Driven Machine Learning System for Driving Style Recognition and Personalized Recommendations," Aug. 2025, [Online]. Available: http://arxiv.org/abs/2509.00802

[61] M. A. Mersha *et al.*, "Explainable AI: XAI-Guided Context-Aware Data Augmentation," Jun. 2025, doi: 10.1016/j.eswa.2025.128364.

[62] S. QU, Z. Gao, X. Wu, and Y. Qiu, "Multi-Attention Fusion Drowsy Driving Detection Model," Dec. 2023, [Online]. Available: http://arxiv.org/abs/2312.17052

[63] Kaggle Inc., "Kaggle Kernels." Accessed: Jan. 24, 2026. [Online]. Available: https://www.kaggle.com/code

[64] M. Abadi *et al.*, “TensorFlow: A system for large-scale machine learning,” May 2016, [Online]. Available: http://arxiv.org/abs/1605.08695

[65] C. Lugaresi *et al.*, “MediaPipe: A Framework for Building Perception Pipelines,” Jun. 2019, [Online]. Available: http://arxiv.org/abs/1906.08172

[66] R. Meudec, “tf-explain: Interpretability Methods for TensorFlow 2.0.” Accessed: Jan. 24, 2026. [Online]. Available: https://github.com/sicara/tf-explain

[67] Rakibul Ecer UET, “Drowsiness Prediction Dataset.” Accessed: Jan. 23, 2026. [Online]. Available: https://www.kaggle.com/datasets/rakibuleceruet/drowsiness-prediction-dataset

[68] R. Ghoddoosian, M. Galib, and V. Athitsos, “A Realistic Dataset and Baseline Temporal Model for Early Drowsiness Detection,” Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.07312

[69] S. Abtahi, M. Omidyeganeh, S. Shirmohammadi, and B. Hariri, “YawDD: A yawning detection dataset,” in *Proceedings of the 5th ACM Multimedia Systems Conference, MMSys 2014*, Association for Computing Machinery, Mar. 2014, pp. 24–28. doi: 10.1145/2557642.2563678.

[70] A. V Deshpande *et al.*, “A Hybrid Approach to Drowsiness Detection Using MediaPipe and YOLOv5,” *IJIRT 170176 INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY*, vol. 3556, 2024.

[71] R. C. Gonzalez and R. E. Woods, “Gaussian smoothing filter,” in *Digital Image Processing*, 4th ed., New York: Pearson, 2018.

[72] M. Fayyaz *et al.*, “Adaptive Token Sampling For Efficient Vision Transformers,” Jul. 2022, [Online]. Available: http://arxiv.org/abs/2111.15667

[73] H. Chefer, S. Gur, and L. Wolf, “Transformer Interpretability Beyond Attention Visualization,” Apr. 2021, [Online]. Available: http://arxiv.org/abs/2012.09838

Anexos

Anexo 1:

Notebook de entrenamiento Modelo Optimizado: https://github.com/Masciel-Sevilla/Tesis-AplicacionTecnicasXAI-Somnolencia/blob/main/entrenamiento_modeloNuevo.ipynb

Anexo 2:

Notebook de aplicación de métricas XAI Modelo Optimizado: https://github.com/Masciel-Sevilla/Tesis-AplicacionTecnicasXAI-Somnolencia/blob/main/metricasxai-nuevo-v3.ipynb

Anexo 3:

Notebook de aplicación de métricas XAI Modelo Base: https://github.com/Masciel-Sevilla/Tesis-AplicacionTecnicasXAI-Somnolencia/blob/main/metricasxai_modeloBase.ipynb

Anexo 4:

Notebook de aplicación de técnicas XAI para ambos modelos: https://www.kaggle.com/code/mascielsevilla/t-cnicasxai-ambosmodelos

Anexo 5:

Dataset con el resultado de las técnicas XAI aplicadas a los 5 casos representativos:

https://www.kaggle.com/datasets/mascielsevilla/resultadosxai-5casos

Anexo 6:

Notebook de aplicación de técnicas XAI en escenarios externos no entrenados (variaciones de luz, uso de gafas y lentes):

https://www.kaggle.com/code/mascielsevilla/casosextras/edit

Anexo 7:

Dataset con el resultado de las técnicas XAI aplicadas a los casos extras):

https://www.kaggle.com/datasets/mascielsevilla/resultadosxai-casosextras