

**DISEÑO DE TÉCNICA DE APRENDIZAJE PROFUNDO  
IMPLEMENTANDO INTELIGENCIA ARTIFICIAL EXPLICABLE A VISIÓN  
ARTIFICIAL PARA LA DETECCIÓN TEMPRANA DEL DETERIORO  
COGNITIVO USANDO LA FIGURA COMPLEJA DE REY**

JUAN PABLO BERNAL GARNICA

BIOINGENIERÍA

UNIVERSIDAD EL BOSQUE

FACULTAD DE INGENIERÍA

2025 - 2

**DISEÑO DE TÉCNICA DE APRENDIZAJE PROFUNDO  
IMPLEMENTANDO INTELIGENCIA ARTIFICIAL EXPLICABLE A VISIÓN  
ARTIFICIAL PARA LA DETECCIÓN TEMPRANA DEL DETERIORO  
COGNITIVO USANDO LA FIGURA COMPLEJA DE REY**

JUAN PABLO BERNAL GARNICA

BIOINGENIERÍA

DIRECTOR DE TESIS:

OSCAR MAURICIO ARIAS

UNIVERSIDAD EL BOSQUE

FACULTAD DE INGENIERÍA

2025 - 2

## RESUMEN

El proyecto propone el desarrollo de un sistema automatizado a manera de herramienta como ayuda en la detección temprana del deterioro cognitivo mediante el análisis de la prueba de Test de Figura Compleja de Rey (TFCR) utilizando técnicas de aprendizaje profundo e Inteligencia Artificial Explicable (XAI por su sigla en inglés). La iniciativa surge ante la creciente prevalencia del deterioro cognitivo en adultos mayores y las limitaciones de las evaluaciones neuropsicológicas tradicionales, que suelen ser subjetivas, demandantes en tiempo y dependientes de personal especializado.

El sistema utilizará una base de datos de 1171 imágenes de pacientes proporcionadas por el Instituto de Neurociencias de la Universidad El Bosque en conjunto con la Fundación Universitaria de Ciencias de la Salud (FUCS). Se plantea analizar el dataset, diseñar e implementar arquitecturas de redes neuronales convolucionales y modelos XAI, y evaluar su desempeño e interpretabilidad clínica, se plantea utilizar modelos de XAI con la finalidad de pulir los modelos de redes neuronales presentados y así mejorar el diseño y la fiabilidad de la herramienta.

El objetivo final fue obtener una herramienta confiable, objetiva y clínicamente interpretable, capaz de optimizar el proceso diagnóstico, facilitar la labor del personal de salud y contribuir a intervenciones tempranas que mejoren la calidad de vida de los pacientes. Los resultados alcanzados demostraron que la integración de Inteligencia Artificial Explicable (XAI) no solo permitió auditar y justificar las decisiones del modelo, sino que también actuó como eje metodológico para su refinamiento, orientando ajustes en arquitectura, aumentación y calibración.

## **ABSTRACT**

The project proposes the development of an automated system as a tool to aid in the early detection of cognitive impairment through the analysis of the Rey Complex Figure Test (RCFT) using deep learning techniques and explainable artificial intelligence (XAI). The initiative arises in response to the growing prevalence of cognitive impairment in older adults and the limitations of traditional neuropsychological assessments, which are often subjective, time-consuming, and dependent on specialized personnel.

The system will use a database of approximately 1,200 patient images provided by the Institute of Neurosciences at El Bosque University in conjunction with the University Foundation for Health Sciences (FUCS). The plan is to analyze the dataset, design and implement convolutional neural network architectures and XAI models, and evaluate their performance and clinical interpretability. The use of XAI models is proposed in order to refine the neural network models presented and thus improve the design and reliability of the tool.

The goal was to obtain a reliable, objective, and clinically interpretable tool capable of optimizing the diagnostic process, facilitating the work of healthcare personnel, and contributing to early interventions that improve patients' quality of life. The results demonstrated that the integration of Explainable Artificial Intelligence (XAI) not only allowed the model's decisions to be audited and justified but also acted as a methodological axis for its refinement, guiding adjustments in architecture, augmentation, and calibration.

## **AGRADECIMIENTOS**

Deseo expresar mi más profundo agradecimiento a mis padres, quienes, con su apoyo incondicional, paciencia y motivación constante hicieron posible la culminación de este proyecto. Su confianza en mis capacidades fue la base que me impulsó a superar cada desafío. Extiendo también mi gratitud a mi tutor de tesis, por su guía experta, orientación metodológica y compromiso en cada etapa del desarrollo, aportando claridad y rigor científico al trabajo. A mis compañeros, gracias por su colaboración, retroalimentación y disposición para compartir conocimientos, lo que enriqueció significativamente el proceso investigativo. Cada aporte, consejo y palabra de aliento contribuyó a que este proyecto se consolidara como una experiencia académica y personal invaluable.

## TABLA DE CONTENIDO

RESUMEN .....	2
ABSTRACT .....	3
AGRADECIMIENTOS .....	4
TABLA DE CONTENIDO .....	5
TABLA DE IMÁGENES .....	7
TABLAS .....	8
INTRODUCCIÓN .....	9
PLANTEAMIENTO DEL PROBLEMA .....	13
MARCO DE REFERENCIA .....	16
Marco teórico .....	16
Deterioro Cognitivo .....	16
Deterioro Cognitivo Leve (DCL) .....	16
Demencia .....	17
Test de la Figura Compleja de Rey (TFCR) .....	18
Evaluación Neuropsicológica Tradicional .....	18
Visión Artificial en Salud .....	19
Redes Neuronales Convolucionales (CNN) .....	19
Clasificación por Percentiles Clínicos .....	20
Normalización .....	20
Eliminación de Ruido .....	20
Técnicas de Data Augmentation .....	21
F1 (Métrica de Desempeño) .....	21
Python .....	22
MATLAB .....	22
Inteligencia Artificial Explicable (XAI) .....	23
Marco normativo .....	25
ESTADO DEL ARTE .....	28
Automating Rey Complex Figure Test scoring using a deep learning-based approach: a potential large-scale screening tool for cognitive decline. ....	28
Sistema de apoyo diagnóstico del deterioro cognitivo mediante el Test de la Figura Compleja de Rey y redes neuronales convolucionales .....	29

A deep learning approach for automated scoring of the Rey–Osterrieth complex figure .....	30
A benchmark for Rey-Osterrieth complex figure test automatic scoring.....	31
Automated scoring of the Rey-Osterrieth Complex Figure Test using a deep-learning algorithm .....	32
OBJETIVOS.....	34
Objetivo General.....	34
Objetivos Específicos .....	34
LEVANTAMIENTO DE REQUERIMIENTOS .....	35
Requerimientos funcionales: .....	35
Requerimientos de calidad: .....	35
Requerimientos de restricción: .....	36
METODOLOGÍA.....	37
Fase 1: Análisis y preparación del conjunto de datos.....	39
Fase 2: Diseño de la red neuronal e implementación de arquitecturas XAI .....	48
Integración de técnicas de inteligencia artificial explicable (XAI) como Grad-CAM, LIME o SHAP para generar visualizaciones interpretables.....	69
Fase 3: Evaluación del desempeño de los modelos.....	75
Fase 4: Evaluación de la interpretabilidad clínica.....	81
Ventajas.....	85
Limitaciones .....	86
RESULTADOS.....	88
CONCLUSIONES.....	113
Conclusiones comparativas con el trabajo previo .....	115
DISCUSION.....	120
SUGERENCIAS.....	123
Anexos.....	126
Anexo NO.1. Códigos Fase 1. ....	126
Anexo NO. 2. Segundo modelo de red neuronal. Fase 2.....	126
Anexo NO. 3. Tercer modelo de red neuronal. Fase 3. ....	126
Anexo NO. 4. XAI.py.....	126
Anexo NO. 5. Código para la clasificación de imágenes por carpetas. Link de acceso a la carpeta en donde se encuentran los archivos, programas, base de datos y bancos de pruebas de la tesis. <a href="https://unbosqueeducos-my.sharepoint.com/:f/g/personal/jbernalg_unbosque_edu_co/EqMLi3GPb1JKhMtXuB0-nqYBvaN2su_I2xO2UT6i52jQsQ?e=ZJHgB3">https://unbosqueeducos-my.sharepoint.com/:f/g/personal/jbernalg_unbosque_edu_co/EqMLi3GPb1JKhMtXuB0-nqYBvaN2su_I2xO2UT6i52jQsQ?e=ZJHgB3</a> .....	126
BIBLIOGRAFÍA.....	127

## TABLA DE IMÁGENES

Ilustración 1. Arbol del problema (Autoia propia) .....	15
Ilustración 2. Diagrama de flujo .....	37
Ilustración 3. TCFR ejemplo Tomada de la base de datos de la Universidad el Bosque .....	40
Ilustración 4. TCFR ejemplo Tomada de la base de datos de la Universidad el Bosque .....	41
Ilustración 5. TCFR ejemplo Tomada de la base de datos de la Universidad el Bosque .....	42
Ilustración 6. Números utilizados para el entrenamiento. (Fuente. Google imágenes.2025).....	60
Ilustración 7. Pérdida y Precisión durante el entrenamiento, autoría propia.....	88
Ilustración 8. Resultados gráficos del entrenamiento mostrando pérdida y precisión durante las fases por época. (Autoría propia). .....	89
Ilustración 9. Pérdida por época y precisión por entrenamiento y prueba. (Autoría propia) .....	92
Ilustración 10. Matriz de confusión para la prueba (Autoría propia) .....	93
Ilustración 11. Matriz de confusión de pruebas calibrada (Autoría propia) .....	94
Ilustración 12. Mapa de calor Grad-CAM de la imagen 1168 (Autoría propia) .....	95
Ilustración 13. LIME realizado por la XAI para la imagen de prueba 1168 (Autoría propia) .....	96
Ilustración 14. Grad-CAM realizado por la XAI para la imagen de prueba numero 1172 (Autoría propia) .....	97
Ilustración 15. LIME realizado por la XAI para la imagen de prueba 1172 (Autoría propia) .....	98
Ilustración 16. Matriz de confusión de pruebas modelo Tercera Fase (Autoría propia) .....	105
Ilustración 17. Matriz de confusión de pruebas calibrada modelo Tercera Fase (Autoría propia).....	106
Ilustración 18. Grad-CAM realizado por la XAI para la imagen de prueba numero 1170 para el.....	108
Ilustración 19. LIME realizado por la XAI para la imagen de prueba 1170 para el entrenamiento de Tercera Fase. (Autoría propia).....	110

## TABLAS

Tabla 1. Comparativa de técnicas de balanceo .....	45
Tabla 2. Tabla comparativa de modelos de análisis de imagen. ....	50
Tabla 3. Comparación de redes neuronales .....	52
Tabla 4. Comparación de librerías para preparación de la red neuronal .....	57
Tabla 5. Resumen de entrenamiento por épocas utilizando el modelo inicial. (Autoría propia).....	64
Tabla 6. Comparativa de XAI disponibles.....	69
Tabla 7. Justificación de selección de Grad-CAM y LIME .....	74
Tabla 8. Tabla de comparación para la selección del early stopper.....	79
Tabla 9. Hallazgos de XAI.py para tomar acciones e impacto (Autoría propia).....	99

## INTRODUCCIÓN

El proyecto se desarrolló con el propósito de diseñar una técnica de aprendizaje profundo que incorporara principios de Inteligencia Artificial Explicable (XAI) aplicada a visión artificial para la detección temprana del deterioro cognitivo mediante el análisis del Test de la Figura Compleja de Rey (TFCR). Esta iniciativa surgió ante la creciente prevalencia del deterioro cognitivo en adultos mayores y las limitaciones de las evaluaciones neuropsicológicas tradicionales, que suelen ser subjetivas, demandantes en tiempo y dependientes de personal especializado.

Durante el trabajo, se abordó la construcción de un sistema capaz de procesar imágenes del TFCR, clasificarlas en categorías clínicas relevantes y ofrecer explicaciones interpretables sobre las decisiones del modelo. Para ello, se utilizó un conjunto de datos compuesto por 1.169 imágenes anonimizadas, que fueron sometidas a procesos de depuración, normalización y balanceo con el fin de garantizar la calidad y representatividad del material. Posteriormente, se implementaron arquitecturas de redes neuronales convolucionales (CNN), seleccionando ResNet-18 como modelo base mediante transfer learning, y se aplicaron técnicas de regularización y calibración para mejorar la estabilidad del aprendizaje.

Un aspecto central del proyecto fue la integración de XAI como componente metodológico, no solo para auditar las decisiones del modelo, sino también para guiar ajustes orientados a reducir sesgos y mejorar la generalización. Herramientas como Grad-CAM y LIME permitieron visualizar las regiones de la imagen que influyeron en la clasificación, revelando inicialmente una dependencia excesiva en rasgos locales. Estas evidencias condujeron a la aplicación de estrategias correctivas, regularización focalizada y fine-tuning selectivo de capas profundas, lo que se tradujo en mejoras sustanciales en las métricas de validación y en la coherencia espacial de los mapas de activación.

En síntesis, el desarrollo del proyecto demostró que la explicabilidad no debe considerarse un complemento opcional, sino un pilar fundamental en el diseño de sistemas de IA aplicados a salud. La capacidad de interpretar y justificar las predicciones incrementó la confianza, habilitó mecanismos de control de calidad y sentó las bases para la adopción ética y regulatoria de la tecnología. Este enfoque confirmó que la XAI actúa como catalizador del refinamiento técnico, guiando la evolución del modelo hacia configuraciones más robustas, auditables y clínicamente útiles.

## **JUSTIFICACIÓN**

La detección temprana del deterioro cognitivo representa un punto crucial para mejorar la calidad de vida de los pacientes, ya que permite intervenir antes de que la progresión clínica se torne irreversible. Estudios han demostrado que identificar los primeros signos de deterioro permite aplicar tratamientos farmacológicos y no farmacológicos que pueden disminuir la velocidad de avance hacia demencia (Petersen, 2018). Este tipo de intervención temprana favorece la preservación de la funcionalidad diaria, prolongando la autonomía personal y retrasando la necesidad de cuidados externos, como lo puede ser el contratar personal especializado para el cuidado de un paciente con deterioro cognitivo desarrollado y sin una red de apoyo inmediata disponible.

Dado lo anterior, en el ámbito clínico, el tener acceso temprano a información que pueda colaborar con la detección temprana del deterioro cognitivo, permite también el acceso a programas como la estimulación cognitiva, la actividad física y el control de factores cardiovasculares ya utilizados en etapas preclínicas o con deterioro cognitivo leve (Ngandu et, 2015).

Dada la importancia de la detección temprana del deterioro cognitivo en pacientes y sobre todo en pacientes con edad avanzada, es importante poder proveer herramientas que ayuden a facilitar el trabajo del personal de la salud en pro de una estrategia para el tratamiento en prevención y control del deterioro cognitivo.

En la investigación previa desarrollada por Cruz (2024) se propuso un sistema de detección temprana de deterioro cognitivo basado en el análisis automatizado del TFCR, mediante visión artificial. Aunque esta

propuesta sentó las bases para una aproximación computacional al diagnóstico, es necesario continuar explorando diferentes métodos haciendo uso de diferentes arquitecturas de redes neuronales para la construcción de una red neuronal capaz de brindar una herramienta que pueda ser usada por el personal médico y brinde apoyo en la decisión del personal de la salud para remitir al paciente en pro del diagnóstico oportuno del deterioro cognitivo. Por lo que el traer a la mesa diferentes modelos permitirá a largo plazo obtener un camino más claro sobre de qué manera se puede guiar una investigación de esta índole y bajo qué arquitecturas o metodologías se puede obtener un mejor resultado en el sistema de detección temprana para el deterioro cognitivo.

La propuesta contempla el análisis detallado de aproximadamente 2000 imágenes de pacientes adultos proporcionadas por el Instituto de Neurociencias de la Universidad El Bosque. Se plantea un proyecto mediante el cual se desarrolle una nueva arquitectura para el sistema de detección temprano de deterioro cognitivo mediante tecnologías más recientes e implementando técnicas de inteligencia artificial explicable (XAI). con la finalidad de buscar una comparación de los resultados obtenidos en el este desarrollo y el presentado previamente por el estudiante Camilo Cruz en la Universidad el Bosque haciendo uso de la misma base de datos, para plantar los cimientos de una investigación más robusta y enfocada en su posible uso por parte del personal de salud como herramienta que ayude a tratamientos de deterioro cognitivo basado en su detección temprana.

## PLANTEAMIENTO DEL PROBLEMA

El deterioro cognitivo en adultos mayores representa un desafío creciente para los sistemas de salud pública, según el Programa Iberoamericano de Cooperación sobre la Situación de las Personas Adultas Mayores (2023) esta condición tiene una prevalencia que oscila entre el 10% y el 20% en adultos mayores de 65 años. Dada su incidencia progresiva y el impacto significativo que tiene sobre la calidad de vida de los pacientes y su entorno. En este contexto, la detección temprana del deterioro cognitivo es fundamental para implementar estrategias de intervención oportunas y eficaces. Sin embargo, los métodos actuales de evaluación diagnóstica suelen depender en gran medida del juicio clínico y de pruebas neuropsicológicas cuya interpretación requiere personal especializado, lo cual limita su alcance y eficiencia en escenarios con alta demanda o pocos recursos.

Tal y como lo expone Howieson (2019) las pruebas de neuropsicología que se utilizan en la actualidad presentan limitaciones significativas ya que muchas requieren la interpretación por parte de profesionales altamente capacitados y no siempre reflejan adecuadamente las habilidades cognitivas en contextos del mundo real. Restringiendo la capacidad de aplicarlas en ámbitos con recursos limitados o con flujos de pacientes elevados.

En Colombia, donde la atención en salud mental enfrenta limitaciones en cobertura, accesibilidad y recursos humanos; este tipo de pruebas no automatizadas se convierte en una barrera adicional para lograr un diagnóstico oportuno, eficaz y objetivo. Esto se agrava en un contexto de envejecimiento poblacional acelerado, en el que cada vez más personas adultas mayores requieren valoración neuropsicológica, aumentando la necesidad de herramientas diagnósticas eficientes, estandarizadas y accesibles (Ardila, Rosselli, & Puente, 2021).

De acuerdo con Guerrero (2023), se habla de envejecimiento poblacional a un ritmo que se proyecta puede alcanzar el 25% para el 2050 en cuanto a adultos mayores de 65 años. También se analiza cómo en el estudio que se realizó, la edad media fue de 70.82 años y se obtuvo una prevalencia de deterioro cognitivo leve o sin demencia del 8.9%, mientras que la prevalencia para el deterioro cognitivo con demencia fue de 10.8% mostrando así un incremento en la prevalencia de deterioro cognitivo con demencia para el tamizaje que se realizó en dicho proyecto. Dada la tendencia a una población cada vez más vieja como se proyecta para el 2050 y la prevalencia de demencia en esta población adulta superior a los 65 años, es necesario así mismo tener herramientas suficientes de detección de deterioro cognitivo.

Existen diversas formas de ayudar al personal de salud en la detección temprana del deterioro cognitivo en las diferentes fases que este se puede llegar a encontrar, una de ellas es el Test de la Figura Compleja de Rey (TFCR), que es una herramienta que es diseñada para la evaluación de las habilidades visoconstructivas y la memoria visual. De acuerdo con Pontón (1996) la habilidad visoconstructiva puede ayudar como medida para: Enfermedades cerebrovasculares, Lesiones cerebrales traumáticas, Alzheimer, Parkinson, epilepsia en el lóbulo temporal. Así como la memoria visual evidencia sensibilidad a los déficits en lesiones como: Lóbulo frontal, Hemisferio derecho, Cerebrales traumáticas, Enfermedad del Parkinson y de Huntington.

El TCFR se compone de una figura geométrica compleja la cual supone una fácil realización gráfica. Esta plantea una evaluación neuropsicológica que provee información sobre el posible estado de deterioro cognitivo en una persona, de acuerdo con Rey (1997), la subjetividad en la puntuación se debe en parte a la necesidad de atención por parte del examinador en las diferentes fases de la prueba, al tratarse de un test cuyo puntaje se evalúa sobre la precisión y posición de diferentes puntos sobre una

imagen clave, al no haber un criterio numérico estandarizado sobre la manera cómo se evaluará cada uno de los puntos que componen la TCFR, este está sujeto a la interpretación del profesional encargado de la calificación del test realizado, al ser un proceso evaluado manualmente también compone limitaciones en términos de tiempo si se tiene un aforo alto en cuanto a pacientes que requieran de esta evaluación.

Estas limitaciones evidencian la urgencia de desarrollar sistemas automatizados que permitan la calificación objetiva y rápida del TFCR, lo que podría optimizar los tiempos de diagnóstico por medio de remisión para su detección y reducir la carga de trabajo del personal clínico y mejorar la precisión en la detección temprana del deterioro cognitivo.

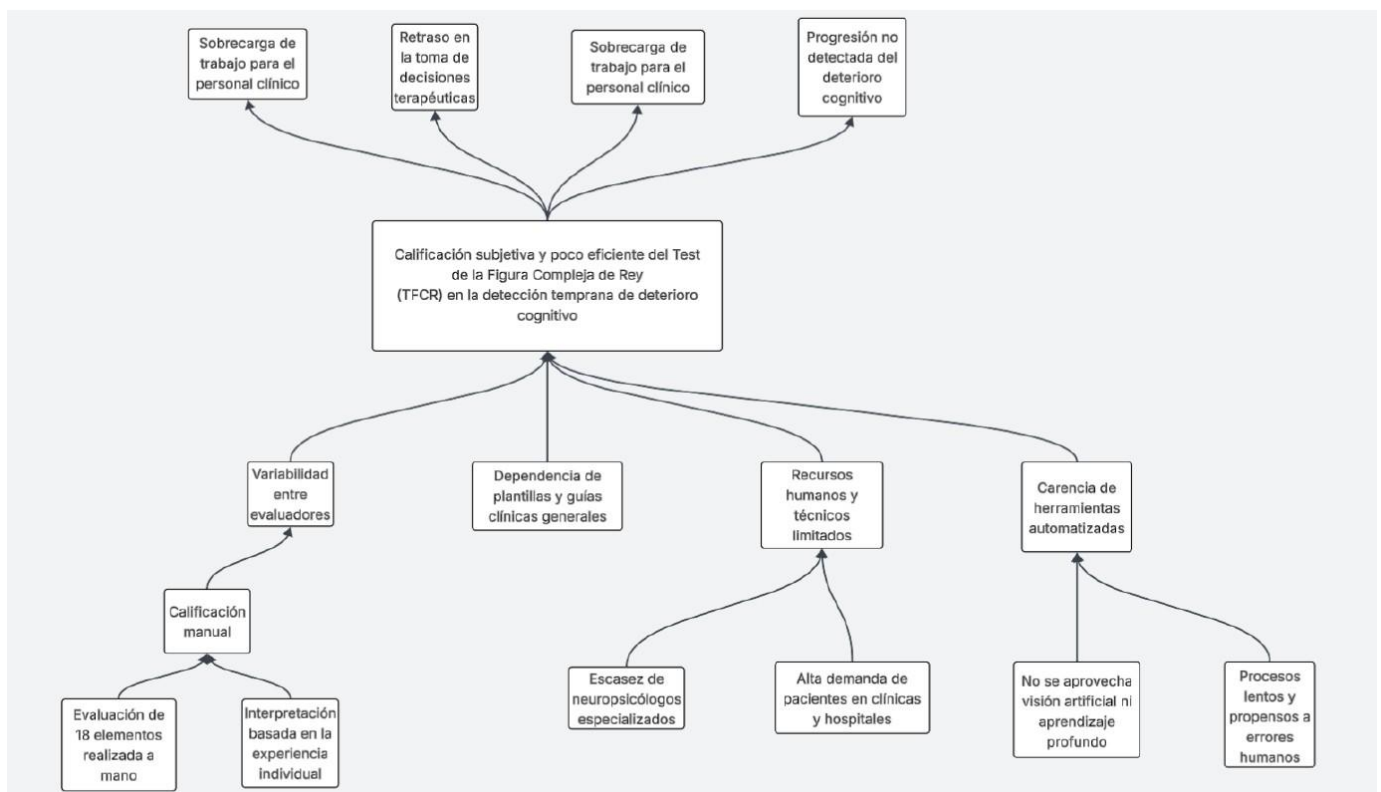


Ilustración 1. Arbol del problema (Autoia propia)

## **MARCO DE REFERENCIA**

### **Marco teórico**

#### **Deterioro Cognitivo**

El deterioro cognitivo es una condición en la cual el paciente se caracteriza por tener una disminución en las funciones mentales superiores, entre las cuales se pueden encontrar la memoria, la atención, el lenguaje y la capacidad de planificar o resolver problemas. El deterioro cognitivo puede ser definido como deterioro cognitivo leve (DCL) o Demencia dependiendo del estado de la persona que lo sufra, el DCL no interfiere en las actividades diarias, aunque puede representar una etapa intermedia hacia la demencia. (Petersen, 2014).

El deterioro cognitivo se entendió como un declive medible de funciones mentales superiores memoria, atención, función ejecutiva y habilidades visoconstructivas que progresó desde variaciones compatibles con la edad hasta cuadros clínicos que afectaron la vida diaria. En el proyecto, esta noción operativa permitió traducir un constructo clínico a categorías observables en la imagen del TFCR, conectando errores de organización espacial, omisiones y distorsiones de proporciones con hipótesis de deterioro. Esta traducción guió tanto el etiquetado de clases como la elección de métricas sensibles al costo clínico de los falsos negativos, y justificó la comparación sistemática entre decisiones del modelo y explicaciones XAI para evitar que correlaciones espurias con el fondo sustituyeran evidencia neuropsicológica genuina.

#### **Deterioro Cognitivo Leve (DCL)**

El Deterioro Cognitivo Leve (DCL) es una condición donde las habilidades mentales disminuyen un poco más de lo que se esperaría según la edad. Sin embargo, esto no afecta mucho las actividades diarias.

Se ve como una etapa intermedia entre el envejecimiento normal y la demencia, especialmente la enfermedad de Alzheimer (Petersen et al., 2014).

El DCL se abordó como un estado intermedio que implicó que los signos fueran sutiles y que la herramienta priorizara sensibilidad sin disparar falsos positivos. Su inclusión forzó al pipeline a detectar patrones finos p. ej., pérdida de jerarquía en el dibujo, fragmentación o leve desalineación y por tanto a diseñar aumentación y regularización que fomentaran atención global y no solo señales burdas. También motivó la calibración de probabilidades y el uso de umbrales de confianza, de modo que predicciones inciertas pudieran derivarse a revisión humana, alineando el sistema con flujos de tamizaje y prevención

## **Demencia**

La demencia es un síndrome neurodegenerativo el cual tiene como característica principal el deterioro progresivo y crónico de funciones cognitivas, entre las cuales se encuentran la memoria, el pensamiento, el lenguaje, la orientación y el juicio. Dado este deterioro hay una interferencia significativa de la autonomía, pues existe una afectación notoria en las actividades cotidianas. (Organización Mundial de la Salud [OMS], 2021) Según la OMS, se estima que más de 55 millones de personas viven con demencia en el mundo y se espera que esta cantidad se triplique para el año 2050. (Alzheimer's Disease International, 2022).

La demencia, como fase avanzada, aportó un ancla que facilitó al modelo aprender el gradiente entre normalidad y DCL. La presencia de rasgos de mayor magnitud (omisiones extensas, desorganización marcada) mejoró la separabilidad en el espacio de características, a la vez que reveló sesgos cuando el clasificador derivaba en exceso hacia la clase dominante. XAI fue decisivo para confirmar que los aciertos en demencia se sustentaran en regiones del dibujo clínicamente pertinentes y no en artefactos del papel o sombra de escaneo, orientando ajustes de preprocesamiento y recorte del campo visual.

## **Test de la Figura Compleja de Rey (TFCR)**

El Test de la Figura Compleja de Rey (TFCR) es una prueba que evalúa cómo las personas manejan las funciones visoespaciales, la memoria visual y las habilidades de planificación. En esta prueba, se pide a los participantes que copien y reproduzcan una figura geométrica compleja. Esto ayuda a entender mejor cómo procesan la información visual y cómo se acuerdan de las cosas (Meyers & Meyers, 1995).

El TFCR se asumió como una ventana gráfica al estado visoconstructivo y a la memoria visual, idónea para visión por computador al tratarse de contornos y proporciones más que de textura o color. Esto justificó la conversión a escala de grises, el redimensionamiento homogéneo y el uso de CNN con sesgo inductivo espacial. La naturaleza modular de la figura (estructura global y subcomponentes) se correspondió con la jerarquía de características de las CNN, mientras que Grad-CAM y LIME permitieron visualizar qué partes de la reproducción influyeron en la decisión, cerrando el círculo entre evidencia clínica y explicación algorítmica.

## **Evaluación Neuropsicológica Tradicional**

La evaluación neuropsicológica tradicional utiliza pruebas estandarizadas para medir diferentes áreas cognitivas. Estas áreas incluyen la atención, la memoria, el lenguaje y la función ejecutiva (Lezak et al., 2012).

Comprender el proceso tradicional de calificación evidenció las limitaciones de subjetividad y demanda de tiempo, lo que definió la propuesta de valor del sistema: estandarizar criterios, acelerar el flujo y ofrecer trazabilidad. Por eso se diseñaron reportes reproducibles (curvas, matrices, CSV por imagen) y módulos XAI

que devolvieron una explicación accesible al clínico (“dónde miró” el modelo), reduciendo la fricción de adopción y permitiendo auditorías y discusiones interdisciplinarias sobre los casos dudosos.

### **Visión Artificial en Salud**

La visión artificial es una parte de la inteligencia artificial que ayuda a los sistemas informáticos a entender y analizar información visual. En el ámbito de la salud, esta tecnología se usa para mejorar el diagnóstico por imágenes, analizar patrones de comportamiento y evaluar pruebas cognitivas (Litjens et al., 2017).

La visión artificial proporcionó el marco para transformar dibujos en descriptores numéricos robustos a variaciones inevitables de captura (iluminación, inclinación, encuadre). En la práctica, esto significó construir un pipeline de preprocesamiento reproducible, definir transformaciones con sentido clínico (rotaciones leves, escalado, blur moderado) y mantener consistencia entre entrenamiento e inferencia. Integrada con XAI, la visión por computador facilitó verificar que el modelo atendiera al objeto clínico el dibujo y no a su contexto, evitando atajos estadísticos que comprometerían la validez.

### **Redes Neuronales Convolucionales (CNN)**

Las redes neuronales convolucionales (CNN) son un tipo de red neuronal diseñada para procesar datos que tienen una estructura de cuadrícula, como las imágenes. Estas redes son muy efectivas para clasificar y reconocer patrones visuales (LeCun et al., 2015).

Las CNN aportaron el sesgo inductivo correcto para dibujos de línea: detectaron bordes, esquinas y composiciones espaciales, elementos que subyacen a la figura de Rey. El uso de transfer learning con ResNet-18 permitió reutilizar filtros genéricos y ajustar solo las capas finales, aprovechando un dataset moderado sin incurrir en altos costos de cómputo. Además, su compatibilidad con métodos de gradientes

posibilitó explicaciones fieles (Grad-CAM), requisito crítico para presentar evidencia interpretable ante personal de salud.

### **Clasificación por Percentiles Clínicos**

La clasificación por percentiles clínicos es una forma útil de comparar resultados con datos de una población de referencia. Esto nos ayuda a ver si un valor es normal o si es inusual (Weiss & Saklofske, 2020).

### **Normalización**

La normalización es una técnica que ajusta diferentes valores a una escala común. Esto hace que sea más fácil comparar variables y mejora el funcionamiento de los algoritmos, especialmente en modelos de aprendizaje automático (Han et al., 2011).

La normalización controló tres fuentes de variabilidad no informativa: tamaño, rango de intensidades y estadística de canales. Redimensionar a  $224 \times 224$  y normalizar con parámetros de ImageNet permitió heredar pesos preentrenados y estabilizar la optimización; al homogeneizar entradas, el modelo dejó de “aprender” el dispositivo de captura para enfocarse en estructura y proporción del dibujo. Esto impactó positivamente la convergencia, la reproducibilidad entre corridas y la comparabilidad de métricas a lo largo de las fases.

### **Eliminación de Ruido**

La eliminación de ruido es un proceso que limpia los datos. Se eliminan valores atípicos o distorsiones que pueden afectar el análisis (Gonzalez & Woods, 2018).

La depuración de imágenes filtro a escala de grises, descarte de duplicados y defectuosos protegió al clasificador de correlaciones espurias con sombras, manchas o márgenes de hoja. En términos prácticos, se observó una mayor concentración de mapas Grad-CAM sobre el objeto tras estas limpiezas y recortes, lo que aumentó la alineación clínica de las explicaciones y redujo la sobreconfianza injustificada en patrones del fondo.

### **Técnicas de Data Augmentation**

Las técnicas de data augmentation ayudan a aumentar el tamaño de un conjunto de datos. Se logran mediante transformaciones, como rotar, cambiar la escala o invertir imágenes (Shorten & Khoshgoftaar, 2019).

La aumentación fue la respuesta metodológica al tamaño y desbalance del dataset. Al introducir variaciones geométricas moderadas, Random Erasing/Cutout y, donde fue pertinente, MixUp/CutMix, se forzó al modelo a utilizar múltiples pistas visuales y no un único rasgo local. XAI sirvió como retroalimentación: cuando los mapas mostraban foco excesivo en un detalle interno, se intensificaron transformaciones que promovieron atención distribuida y robustez a orientación/escala, logrando mejoras en validación y menor dispersión fuera de la diagonal en las matrices de confusión.

### **F1 (Métrica de Desempeño)**

La métrica F1 combina precisión y recuperación en un solo valor. Mide el desempeño en clasificación y es especialmente útil cuando hay un desequilibrio entre las clases. Ofrece un buen balance entre falsos positivos y falsos negativos (Sammut & Webb, 2017).

Con clases de frecuencia desigual, la accuracy resultó insuficiente para juzgar valor clínico. El F1 por clase y el macro-F1 aportaron una lectura equilibrada entre precisión y recall, visibilizando si el sistema sacrificaba sensibilidad en DCL o demencia para “ganar” exactitud global. Estas métricas guiaron decisiones como ajustar pesos de clase, revisar el régimen de aumentación, introducir early stopping y evaluar la calibración de probabilidades para umbrales de derivación en escenarios de baja confianza.

## **Python**

Python es un lenguaje de programación muy popular en ciencia de datos y desarrollo de modelos de inteligencia artificial (Van Rossum & Drake, 2009).

Python articuló el ciclo completo de ciencia de datos: ingestión, preprocesamiento, entrenamiento, evaluación, calibración y XAI. Con PyTorch, torchvision y scikit-learn, se implementaron DataLoaders reproducibles, persistencia de artefactos (best/last model, classes.json, temperature.json), curvas y matrices, y se encapsularon entradas para LIME. Esta integración redujo fricción operativa, aceleró iteraciones “hallazgo-acción-verificación” y posibilitó que la trazabilidad fuera una propiedad intrínseca del pipeline.

## **MATLAB**

MATLAB es un entorno de programación numérica que se usa en ingeniería, análisis de señales, imágenes y simulaciones. Su interfaz facilita el trabajo con matrices y visualizaciones (MathWorks, 2022).

Aunque no constituyó el núcleo del entrenamiento, MATLAB se consideró un entorno útil para prototipado y verificación de etapas de filtrado o análisis geométrico, así como para exploraciones rápidas de preprocesamiento. Su presencia en el marco teórico dejó abierta la puerta a interoperabilidad futura (por

ejemplo, análisis de segmentos o secuencias de trazos si se dispusiera de información temporal), sin tensionar el hilo conductor del pipeline principal en Python.

### **Inteligencia Artificial Explicable (XAI)**

La Inteligencia Artificial Explicable (XAI) busca que los modelos sean interpretables por humanos sin perder rendimiento, especialmente en dominios críticos como la medicina (Gunning & Aha, 2019). A diferencia de los modelos de caja negra, como muchas redes neuronales profundas tradicionales, la XAI busca desarrollar algoritmos cuya lógica de decisión pueda ser comprendida, auditada y validada por expertos humanos. Esto resulta especialmente relevante en la detección temprana del deterioro cognitivo, donde las decisiones del sistema automatizado deben ser confiables, interpretables y clínicamente justificables.

En el contexto del análisis del Test de la Figura Compleja de Rey (TFCR), integrar principios de XAI permite no solo mejorar la precisión del modelo, sino también facilitar que los profesionales de la salud comprendan qué aspectos de la imagen (como los elementos mal copiados o la secuencia de trazos) influyeron en la clasificación. De este modo, el sistema no sólo entrega un resultado, sino que ofrece una explicación visual o lógica sobre su diagnóstico, fortaleciendo su utilidad como herramienta de apoyo clínico y reduciendo la resistencia a su adopción en ambientes reales.

Grad-CAM y LIME aportan componentes complementarios. Grad-CAM (Gradient-weighted Class Activation Mapping) aprovecha los gradientes asociados a la clase objetivo en la última capa convolucional para localizar las regiones de la imagen que más contribuyen a la predicción, generando mapas de calor sin necesidad de reentrenamiento (Selvaraju et al., 2017). Sus autores muestran utilidad para diagnosticar modos de fallo, revelar sesgos del conjunto de datos y aumentar la confianza del usuario en estudios con participantes (Selvaraju et al., 2017). Por su parte, LIME (Local Interpretable

Model-agnostic Explanations) explica predicciones individuales mediante perturbaciones locales de la entrada en superpíxeles y el ajuste de un modelo interpretable que estima la contribución de cada región a la salida, al ser agnóstico al modelo, proporciona una validación independiente de los mecanismos internos (Ribeiro et al., 2016). La convergencia de evidencias entre un método fiel al modelo (Grad-CAM) y otro agnóstico y local (LIME) fortalece la validez interna de las interpretaciones (Ribeiro et al., 2016; Selvaraju et al., 2017).

## **Marco normativo**

### Ley 1581 de 2012 – Protección de Datos Personales

Establece disposiciones generales para la protección de datos personales, garantizando el derecho al habeas data.

Aplicación al proyecto:

- Requiere autorización expresa de los pacientes para el uso de imágenes del Test de la Figura Compleja de Rey (TFCR).
- Obliga a implementar medidas de seguridad y confidencialidad en el tratamiento de los datos.
- Las imágenes deben estar anonimizadas si se comparten para entrenamiento del modelo.

### Ley 1438 de 2011 – Reforma del Sistema General de Seguridad Social en Salud

Promueve el fortalecimiento del sistema de salud, incluyendo la prevención y diagnóstico temprano de enfermedades.

Aplicación al proyecto:

- El proyecto se alinea con el principio de prevención en salud pública.
- Contribuye al objetivo de mejorar el acceso y la eficiencia del diagnóstico en poblaciones vulnerables (adultos mayores).

### Ley 23 de 1981 – Normas sobre Ética Médica

Regula la conducta ética del ejercicio médico, especialmente en relación con el respeto por la dignidad del paciente.

Aplicación al proyecto:

- El uso de datos clínicos debe ser aprobado por un comité de ética en investigación.
- Toda intervención diagnóstica asistida por IA debe estar supervisada por un profesional de la salud calificado.

Resolución 8430 de 1993 – Normas científicas, técnicas y administrativas para la investigación en salud  
Norma que regula los principios éticos y técnicos de las investigaciones en salud con seres humanos.

Aplicación al proyecto:

Clasifica este proyecto como investigación con riesgo mínimo, al usar imágenes clínicas anonimizadas.

- Requiere revisión y aprobación por un comité de ética en investigación.
- Se debe contar con consentimiento informado de los participantes o del custodio de los datos.

Ley 1751 de 2015 – Ley Estatutaria de Salud

Reconoce el derecho fundamental a la salud, incluyendo el acceso a servicios de diagnóstico y tratamiento oportuno.

Aplicación al proyecto:

- Favorece la implementación de tecnologías que mejoren la cobertura diagnóstica, como las soluciones basadas en IA.
- Este proyecto contribuye a garantizar ese derecho con herramientas más rápidas y objetivas.

Ley 1341 de 2009 (modificada por Ley 1978 de 2019) – Ley TIC

Fomenta el desarrollo y apropiación de tecnologías de la información y la comunicación en sectores estratégicos.

Aplicación al proyecto:

- Impulsa la adopción de soluciones digitales en el sector salud.
- Brinda soporte legal para el uso de IA como parte del ecosistema TIC nacional.

## ESTADO DEL ARTE

### **Automating Rey Complex Figure Test scoring using a deep learning-based approach: a potential large-scale screening tool for cognitive decline.**

El Test de la Figura Compleja de Rey (TCFR) es una herramienta clave en neuropsicología para evaluar habilidades cognitivas en diversos grupos clínicos, desde jóvenes hasta adultos mayores. Sin embargo, su sistema de calificación complejo puede dar lugar a discrepancias entre evaluadores, lo que complica su uso estandarizado. En este estudio, un equipo de investigación desarrolló un método de puntuación automatizado para el TCFR utilizando inteligencia artificial (IA) y un algoritmo de aprendizaje profundo. Para crear este método, se analizaron 20,040 dibujos del TCFR de 6,680 personas registradas en la cohorte Gwangju Alzheimer 's and Related Dementia (GARD). Se implementó un sistema con la arquitectura DenseNet y se mejoró la calidad de las imágenes utilizadas para entrenar el modelo, lo que garantiza un mejor rendimiento.

Los resultados fueron alentadores: el sistema de IA logró un alto nivel de precisión en la calificación del TCFR, con mínimas diferencias en comparación con las evaluaciones realizadas por psicólogos experimentados. Además, mostró la capacidad de diferenciar entre personas con deterioro cognitivo leve, demencia y aquellas sin alteraciones cognitivas significativas.

En conclusión, los investigadores proponen que este sistema de puntuación basado en IA podría ser una herramienta efectiva para el cribado masivo, facilitando la identificación temprana y precisa de posibles casos de enfermedad de Alzheimer.

Este antecedente constituye un aporte significativo al campo, al validar la viabilidad técnica de emplear redes neuronales profundas para asistir el diagnóstico de deterioro cognitivo. Sin embargo, su aproximación presenta ciertas limitaciones que este proyecto busca abordar. En primer lugar, el modelo de Lee et al. Se centra en una arquitectura específica (DenseNet), sin realizar un análisis comparativo entre distintas alternativas de redes convolucionales que podrían optimizar el desempeño del sistema en diferentes contextos poblacionales.

### **Sistema de apoyo diagnóstico del deterioro cognitivo mediante el Test de la Figura Compleja de Rey y redes neuronales convolucionales**

En el estudio de Maldonado, Salazar, Puente y Ávila (2024) se propone un sistema de apoyo diagnóstico automatizado utilizando aprendizaje profundo para la clasificación de deterioro cognitivo a partir del Test de la Figura Compleja de Rey-Osterrieth (TFCRO), específicamente en su fase de copia inmediata. El sistema está basado en una arquitectura de red neuronal convolucional (CNN) adaptada a partir del modelo MobileNet, implementado en el entorno de Google Colab.

Se utilizó un conjunto de 410 imágenes etiquetadas (257 normales, 153 con diagnóstico de deterioro cognitivo), extraídas de evaluaciones clínicas y del estudio de prevalencia del deterioro cognitivo en Bogotá (Pedraza et al., 2017). Las imágenes fueron preprocesadas mediante redimensionamiento a 600×600 píxeles, conversión a escala de grises y normalización. Debido al desbalance de clases, se aplicaron técnicas de data augmentation (rotación aleatoria) alcanzando una base extendida de 5162 imágenes para la clase demencia y 4770 para la clase normal.

El modelo fue entrenado usando MobileNet adaptado con imágenes reescaladas a  $224 \times 224 \times 3$ , incorporando 30 épocas, una tasa de aprendizaje de 0.001, batch size de 40, función de activación ReLU y regresión logística en la capa de salida. La función de error utilizada fue MSE (Error Cuadrático Medio).

Los resultados mostraron un accuracy global del 84%, con una precisión del 92% para la clase deterioro cognitivo, y del 72% para la clase normal, además de un F1-score de 87% y 79% respectivamente. Estos resultados evidencian la viabilidad de emplear modelos CNN como herramientas de apoyo diagnóstico en contextos clínicos, especialmente en atención primaria, donde la disponibilidad de tiempo y la experticia para aplicar e interpretar pruebas neuropsicológicas suele ser limitada.

Este trabajo destaca el potencial de la IA aplicada directamente a instrumentos gráficos como el TFCRO, diferenciándose de la mayoría de investigaciones previas centradas en escalas cuantitativas o datos clínicos estructurados.

### **A deep learning approach for automated scoring of the Rey–Osterrieth complex figure**

Langer et al. (2024) desarrollaron un sistema automatizado para calificar la Figura Compleja de ReyOsterrieth (ROCF), una prueba neuropsicológica que evalúa la memoria visuoespacial. La calificación manual del ROCF es laboriosa y subjetiva, lo que genera variabilidad. Para resolver esto, los autores crearon un modelo de aprendizaje profundo utilizando más de 20,000 imágenes de ROCF dibujadas a mano de individuos sanos y pacientes con trastornos neurológicos y psiquiátricos.

El sistema utiliza una red neuronal convolucional multi-cabezal que combina regresión y clasificación multi-etiqueta. Para mejorar la precisión y la robustez, se entrenó con aumento de datos y aumento de datos de prueba. Los resultados muestran que el modelo de IA supera a los clínicos y evaluadores humanos en cuanto a rendimiento, lo que demuestra menor parcialidad y mayor precisión en la calificación de los elementos individuales del ROCF. La robustez del modelo se validó en diversas condiciones, grupos demográficos y estados clínicos, lo que sugiere su potencial para una evaluación confiable, objetiva y eficiente del ROCF.

Frente a este panorama, el presente proyecto propone una evolución metodológica: no solo automatizar la calificación del TFCR, sino también comparar diferentes arquitecturas de redes neuronales convolucionales para determinar cuál ofrece el mejor balance entre precisión, eficiencia computacional y aplicabilidad clínica.

### **A benchmark for Rey-Osterrieth complex figure test automatic scoring.**

Guerrero-Martín et al. (2024) presentan un marco de referencia para evaluar automáticamente el Test de la Figura Compleja de Rey-Osterrieth (ROCF). Este marco incluye el primer conjunto de datos abierto de dibujos lineales del ROCF, llamado ROCFD528, y los resultados experimentales de varios modelos modernos de aprendizaje profundo. El objetivo principal es facilitar la comparación justa de los sistemas de visión artificial que automatizan esta tarea neuropsicológica compleja, que carece de puntos de referencia públicos y dificulta el progreso en el campo.

Los autores evaluaron varias redes neuronales convolucionales (CNN) de última generación bajo paradigmas de aprendizaje tradicional y transferencia. Los resultados experimentales cuantitativos mostraron que una CNN diseñada específicamente para bocetos superó a otras arquitecturas de CNN de última generación cuando se dispone de pocos ejemplos.

El estudio destaca el potencial del marco de referencia para desarrollar modelos eficientes y robustos que analicen dibujos lineales y bocetos, no sólo para tareas de clasificación sino también de regresión, dentro del campo más amplio del aprendizaje automático. Además, el conjunto de datos ROCFD528 se presenta como un recurso valioso para investigar la detección temprana del deterioro cognitivo en la población anciana.

#### **Automated scoring of the Rey-Osterrieth Complex Figure Test using a deep-learning algorithm**

Vogt et al. (2019) desarrollaron un algoritmo automatizado para puntuar el Test de la Figura Compleja de Rey-Osterrieth (ROCF) y comprobaron su eficacia comparando los resultados del algoritmo con las puntuaciones de evaluadores humanos. El algoritmo se basó en una cascada de redes neuronales profundas entrenadas con las puntuaciones de evaluadores humanos para identificar los 18 segmentos de la figura y medir el rendimiento del paciente. Los resultados del algoritmo se compararon con los de seis evaluadores expertos para 303 dibujos.

Los investigadores analizaron si la correlación promedio entre las puntuaciones del algoritmo y las de los evaluadores humanos era similar a la correlación promedio entre los evaluadores (con un límite de igualdad  $\Delta r < .05$ ). Se utilizaron pruebas de recuerdo inmediato y diferido; la prueba de copia mostró un fuerte efecto techo. Los resultados mostraron que la correlación media de Pearson entre los evaluadores fue de .94 (DE = 0.01), mientras que la correlación entre el algoritmo y los evaluadores fue de .88 (DE

= 0.02).

Una prueba de equivalencia de pruebas t de dos a uno (TOST) reveló que estas correlaciones no eran estrictamente equivalentes,  $t(5) = 4.02$ ,  $p = .995$ , 95% CI [0.35, 0.52]. Aunque no eran estrictamente equivalentes a las clasificaciones humanas, el rendimiento del algoritmo es alto y se acerca al nivel de confiabilidad encontrado entre los evaluadores humanos. Los autores esperan que la mejora en la detección de segmentos individuales permita que la precisión de la puntuación del algoritmo se iguale a la de los evaluadores humanos. La puntuación algorítmica del ROCF probablemente ahorrará tiempo valioso y conducirá a una mayor estandarización en la práctica clínica.

## **OBJETIVOS**

### **Objetivo General**

Desarrollar un sistema de detección de deterioro cognitivo temprano usando técnicas de procesamiento de imágenes bajo la prueba de la figura Compleja de Rey con una arquitectura de aprendizaje profundo explicable.

### **Objetivos Específicos**

- Analizar el dataset disponible compuesto por imágenes la prueba de la Figura Compleja de Rey, con el fin de sustraer información de la caracterización en pro de usar el mismo modelo o aplicar algún cambio sobre este.
- Diseñar técnicas de aprendizaje neuronal artificial profundo explicable (XAI) para la clasificación interpretada para la detección temprana del deterioro cognitivo utilizando la figura compleja de Rey.
- Evaluar el desempeño de las arquitecturas empleadas en términos de precisión y eficiencia computacional en la clasificación de las imágenes.
- Analizar la interpretabilidad clínica del modelo presentado por Cruz, mediante técnicas de inteligencia artificial explicable, con el propósito de validar su aplicabilidad como herramienta de soporte al diagnóstico por parte del personal de salud.

## LEVANTAMIENTO DE REQUERIMIENTOS

### Requerimientos funcionales:

- El sistema debe otorgar un puntaje para cada uno de los elementos a evaluar.
- El sistema debe clasificar la imagen otorgada por el usuario en el posible nivel de deterioro cognitivo.
- El sistema debe utilizar la información otorgada por el usuario en la entrada de datos para la clasificación de los criterios brindando una herramienta de soporte para la detección de deterioro cognitivo.
- El sistema debe retroalimentarse mediante XAI para el conocimiento en toma de decisiones por el modelo presentado.

### Requerimientos de calidad:

- El sistema debe tener una precisión de 80 - 95%||
- El sistema deberá aplicar una retroalimentación mediante XAI en la toma de decisiones para la clasificación de las imágenes.
- Todos los experimentos y ejecuciones deben ser trazables, de modo que cualquier prueba pueda replicarse y validarse por terceros.

### **Requerimientos de restricción:**

Se utilizará una base de datos otorgada por la Universidad el Bosque con un aproximado de 2000 imágenes y un promedio de 65 años de edad.

- La base de datos se utilizará para fines de entrenamiento de la red neuronal y no estará a disposición del usuario.
- Se realizará el entrenamiento de la red neuronal mediante el uso de la fase de copia de los pacientes únicamente.
- El sistema no hará un análisis entre las dos metodologías presentadas a comparar
- El sistema no usará datos que no se encuentren en el cluster presentado por la Universidad el Bosque.

## METODOLOGÍA

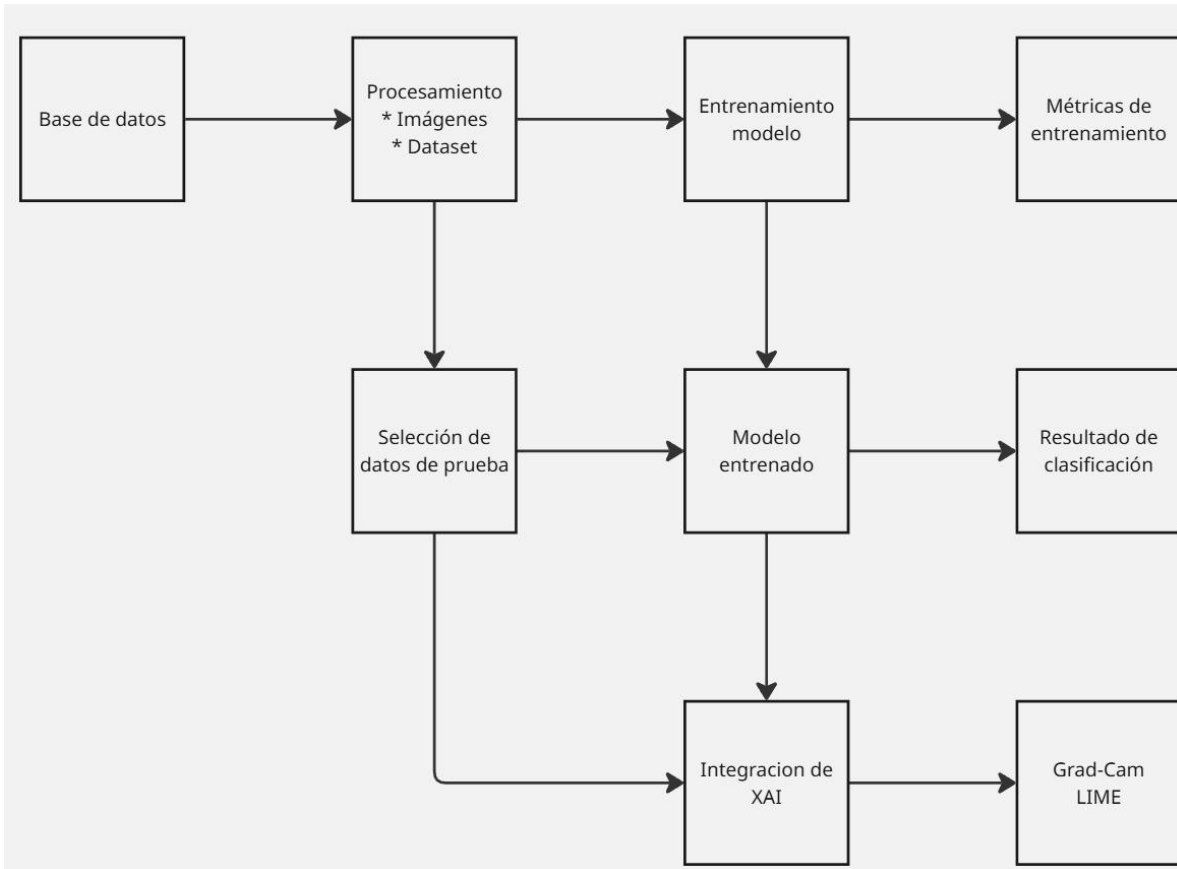


Ilustración 2. Diagrama de flujo

El diagrama presentó, a manera de síntesis, el flujo metodológico que se implementó desde la llegada de los datos hasta la obtención de predicciones explicables. En primer lugar, ingresó la base de datos original con las imágenes del TFCR y sus etiquetas clínicas, las cuales fueron sometidas a un proceso de depuración y estandarización: se eliminaron registros duplicados, incompletos o de baja calidad, se convirtieron las imágenes a escala de grises, se redimensionaron a  $224 \times 224$  píxeles y se normalizaron con los estadísticos compatibles con la arquitectura preentrenada. En esta misma etapa se abordó el desbalance severo entre clases; dado que la clase 1 triplicaba el número de muestras respecto a las otras dos, se optó por un descarte controlado de imágenes de la clase mayoritaria y, de forma complementaria, se aplicó aumentación moderada sobre el conjunto de entrenamiento, con lo cual se consolidó un dataset más equilibrado y representativo para el aprendizaje.

Una vez preparado el conjunto de datos, el pipeline condujo dichas entradas al módulo de entrenamiento del modelo. Allí se ejecutó transferencia de aprendizaje con una CNN preentrenada (ResNet-18), se ajustaron las capas finales y, cuando fue pertinente, se descongelaron capas profundas con tasas de aprendizaje diferenciadas. Durante las épocas de entrenamiento, el código calculó la función de pérdida y las métricas de desempeño sobre entrenamiento y validación, persistió los artefactos del experimento (mejor y último modelo, mapeo de clases) y registró curvas de convergencia. En paralelo, a partir del dataset ya procesado se definió un conjunto de prueba independiente y estratificado, que se conservó intacto para la evaluación final, evitando fugas de información.

Con el modelo entrenado, se procedió a la inferencia sobre el conjunto de prueba. El sistema produjo, por cada imagen, una etiqueta predicha y una probabilidad asociada, y consolidó resultados en reportes reproducibles: exactitud global, métricas por clase (precisión, recall, F1), matriz de confusión y archivos por instancia que documentaron la etiqueta verdadera, la predicción y la confianza. Cuando se requirió, se calibraron las probabilidades mediante escalado de temperatura para alinear la confianza reportada con la frecuencia real de aciertos, habilitando umbrales de decisión y derivación a revisión humana en casos de incertidumbre.

Finalmente, el modelo optimizado y el subconjunto de prueba alimentaron el módulo de Inteligencia Artificial Explicable. En esta fase se integraron Grad-CAM y LIME para generar explicaciones visuales y locales que mostraron qué regiones de cada imagen sustentaron la decisión del clasificador. Las salidas consistieron en mapas de calor superpuestos a las imágenes y representaciones por superpíxeles con las contribuciones positivas, que se almacenaron de forma sistemática para comparación antes y después de ajustes. En conjunto, el flujo aseguró que lo que entró como un banco heterogéneo de imágenes terminó convertido en un sistema capaz de ofrecer predicciones trazables y explicables: se controló el sesgo de clase mediante el balanceo por descarte, se estandarizó la

calidad de las entradas, se entrenó y evaluó el modelo con métricas robustas y, sobre todo, se documentó por qué el modelo decidió lo que decidió, condición clave para su aplicabilidad clínica.

### **Fase 1: Análisis y preparación del conjunto de datos**

Objetivo asociado: Analizar el conjunto de datos disponible.

Actividades:

- Revisión del contenido del dataset del Test de la Figura Compleja de Rey (TFCR) entregado por el Instituto de Neurociencias de la Universidad El Bosque (1172 imágenes).

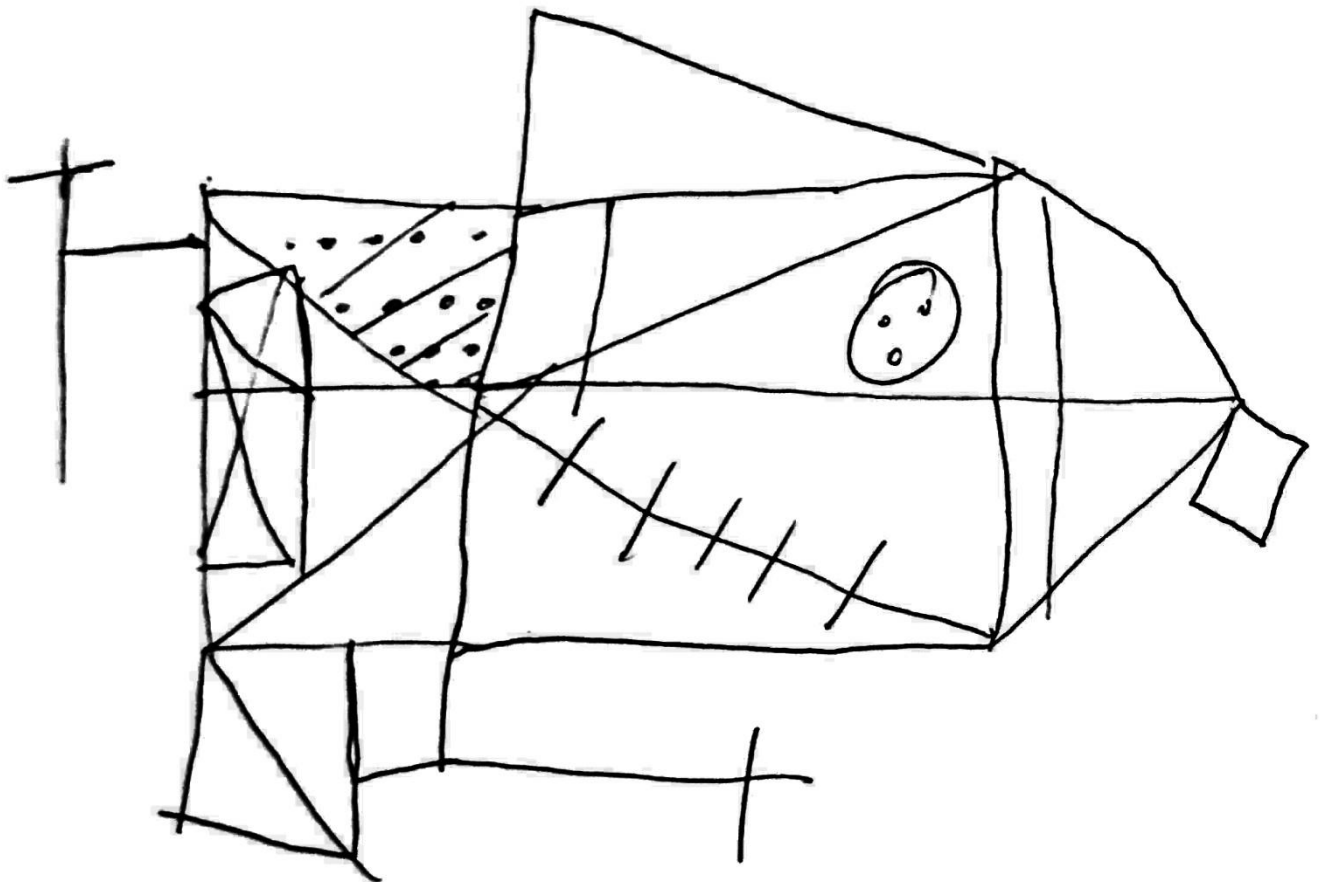
La primera actividad consistirá en la revisión y depuración del dataset del Test de la Figura Compleja de Rey (TFCR) proporcionado por el Instituto de Neurociencias de la Universidad El Bosque. Inicialmente, el conjunto contaba con 1172 imágenes; sin embargo, tras un proceso de filtrado y estandarización, se consolidó una base definitiva de 1172 imágenes en formato blanco y negro. Esta reducción obedeció a la eliminación de registros duplicados, incompletos o de baja calidad, con el fin de asegurar un insumo confiable para el desarrollo del modelo.

El filtrado a escala de grises se implementó con el propósito de optimizar el procesamiento digital de las imágenes. Esta decisión técnica aporta varias ventajas: en primer lugar, permite reducir el ruido y la variabilidad cromática, que no constituye un factor relevante para la interpretación de la prueba (Gonzalez & Woods, 2018); en segundo lugar, simplifica las características visuales de entrada, concentrando la atención del modelo en los trazos y estructuras geométricas esenciales del TFCR, lo que resulta

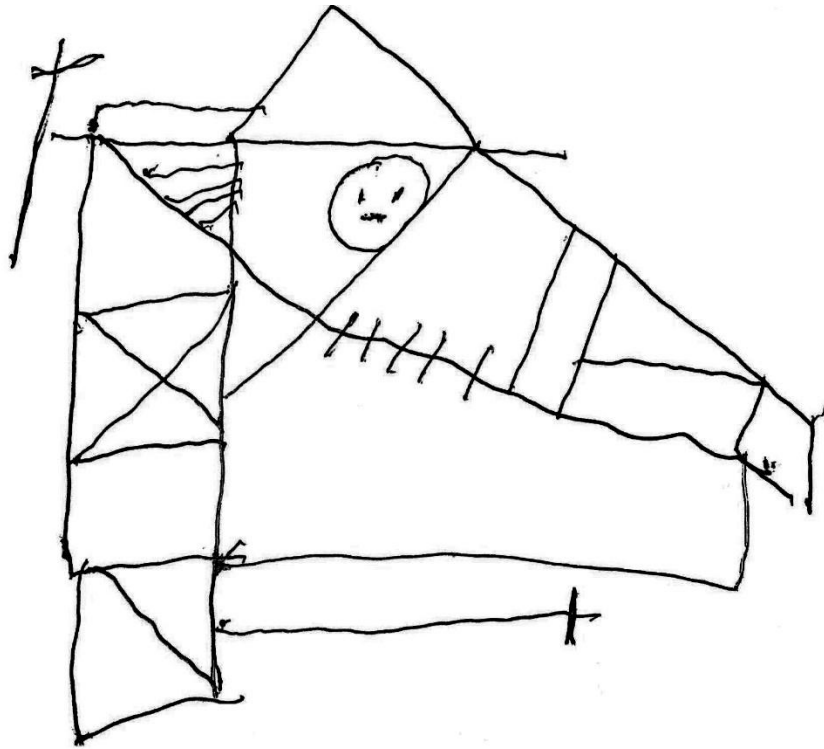
fundamental en tareas de clasificación basadas en contornos y formas (Zhou, Greenspan, & Shen, 2019); y, finalmente, contribuye a disminuir la complejidad computacional, ya que reduce la cantidad de canales de información que deben ser procesados, acelerando así las etapas de entrenamiento y validación (Han, Kamber, & Pei, 2011).

De esta manera, contar con un conjunto homogéneo de imágenes en blanco y negro garantiza no solo una comparación más objetiva y precisa entre las muestras, sino también una mayor eficiencia en el diseño y evaluación de las arquitecturas de aprendizaje profundo propuestas en el proyecto.

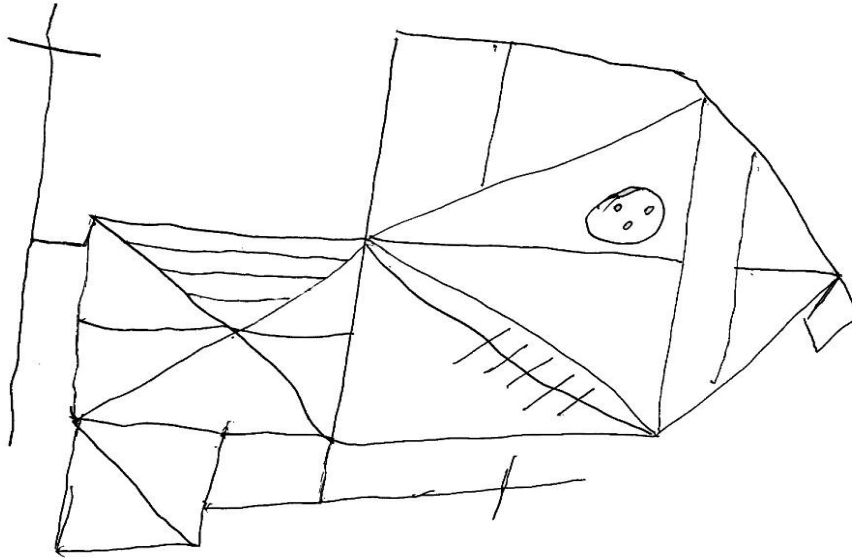
Imágenes filtradas que se usaran en el entrenamiento de la red neuronal convolucional.



*Ilustración 3. TCFR ejemplo Tomada de la base de datos de la Universidad el Bosque*



*Ilustración 4. TCFR ejemplo Tomada de la base de datos de la Universidad el Bosque*



*Ilustración 5. TCFR ejemplo Tomada de la base de datos de la Universidad el Bosque*

- Normalización de las imágenes (resolución, color/escala de grises, formato).

La normalización de las imágenes del dataset fue un paso esencial para garantizar la consistencia y estabilidad del entrenamiento de la red neuronal. Inicialmente, cada imagen fue redimensionada a  $224 \times 224$  píxeles, lo cual permitió unificar la resolución de entrada y hacer compatible el dataset con arquitecturas preentrenadas como ResNet-18, que requieren entradas de ese tamaño (LeCun, Bengio, & Hinton, 2015).

Posteriormente, se realizó la conversión a escala de grises con tres canales (`transforms.Grayscale(num_output_channels=3)`), lo que conserva la información esencial de los trazos y

formas de la Figura Compleja de Rey al tiempo que asegura la compatibilidad con modelos preentrenados en imágenes RGB. Esta estrategia mantiene la simplicidad de la representación en escala de grises, pero la adapta a los requisitos estructurales de las redes convolucionales (Zhou, Greenspan, & Shen, 2019).

El siguiente paso fue la transformación de las imágenes a tensores numéricos (`transforms.ToTensor()`), que normaliza los valores de píxel al rango  $[0,1]$ , reduciendo el impacto de diferencias en escala de intensidad. Finalmente, se aplicó una normalización estandarizada con medias y desviaciones típicas específicas de ImageNet (`Normalize(mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225])`). Esto ajusta cada canal de la imagen para que presente una distribución estadística similar a la de los datos con los que se entrenó originalmente ResNet, favoreciendo la transferencia de conocimiento y la convergencia más rápida y estable del modelo (Han, Kamber, & Pei, 2011; Sammut & Webb, 2017).

En conjunto, este procedimiento asegura que las imágenes ingresadas al modelo estén homogeneizadas en resolución, intensidad y escala estadística, reduciendo la varianza entre muestras y mejorando la precisión en la detección automática de patrones relacionados con el deterioro cognitivo.

- Limpieza y balanceo del dataset si existen clases desproporcionadas.

La etapa de limpieza y balanceo del dataset se llevó a cabo con el propósito de garantizar que el modelo de aprendizaje profundo no incurriera en sesgos derivados de una distribución desigual de clases. Inicialmente, durante la organización de las imágenes en carpetas por categorías (ej. normal, deterioro cognitivo leve, demencia), se verificó la existencia de registros duplicados, incompletos o de baja calidad,

los cuales fueron eliminados para preservar únicamente muestras válidas. Esta depuración permitió consolidar un dataset confiable de 1172 imágenes preprocesadas.

Posteriormente, se abordó el problema del desbalance de clases, común en tareas biomédicas, donde suele existir un número mayor de casos normales frente a casos patológicos (Litjens et al., 2017). Un dataset desbalanceado puede llevar al modelo a aprender patrones sesgados hacia la clase mayoritaria, reduciendo la sensibilidad para detectar casos clínicamente relevantes (He & Garcia, 2009). Para mitigar este efecto, se aplicaron técnicas de data augmentation en el conjunto de entrenamiento, como la inversión horizontal aleatoria (`transforms.RandomHorizontalFlip()` en el código), que permiten aumentar artificialmente la variabilidad de las muestras minoritarias sin alterar la esencia de la información diagnóstica.

Durante la etapa de limpieza y balanceo del dataset se evaluaron diferentes estrategias para mitigar el sesgo derivado de la distribución desigual de clases. Si bien se aplicaron técnicas de data augmentation para incrementar la variabilidad en las clases minoritarias, el desbalance inicial era significativo: la clase 1 (deterioro cognitivo leve) triplicaba el número de imágenes respecto a las clases normal y demencia. Ante esta situación, la técnica más efectiva fue la eliminación controlada de imágenes pertenecientes a la clase mayoritaria, lo que permitió consolidar un conjunto más equilibrado sin comprometer la diversidad diagnóstica. Esta decisión redujo la tendencia del modelo a sobreajustarse a la clase dominante y mejoró la sensibilidad hacia las categorías clínicamente críticas, garantizando que el sistema mantuviera un comportamiento más justo y representativo en la clasificación.

Tabla 1. Comparativa de técnicas de balanceo

<b>Técnica</b>	<b>Descripción</b>	<b>Ventajas</b>	<b>Limitaciones</b>
Data Augmentation	Generación de imágenes sintéticas mediante transformaciones (rotación, flip).	Incrementa variabilidad sin recolectar nuevos datos.	No reduce el desbalance real; riesgo de sobreajuste si se aplica de forma agresiva.
Reponderación de clases	Ajuste de pesos en la función de pérdida para penalizar la clase mayoritaria.	Fácil de implementar; no altera el dataset original.	Puede inducir inestabilidad en el entrenamiento si el desbalance es extremo.
Descartar imágenes	Eliminación controlada de muestras de la clase mayoritaria para equilibrar.	Reduce sesgo de manera directa; mantiene integridad de clases minoritarias.	Disminuye tamaño total del dataset, pero mejora representatividad global.

De manera adicional, la implementación del DataLoader con el parámetro `shuffle=True` en el entrenamiento aseguró que las imágenes fueran presentadas al modelo en un orden aleatorio, contribuyendo a reducir la dependencia de secuencias sesgadas y favoreciendo una mejor generalización (Goodfellow, Bengio, & Courville, 2016). Estas estrategias en conjunto no solo fortalecen la capacidad de aprendizaje del modelo, sino que garantizan un entrenamiento más justo y representativo de todas las clases, mejorando la sensibilidad y especificidad en la clasificación del deterioro cognitivo.

- Generación de nuevas muestras mediante técnicas de data augmentation.

La generación de nuevas muestras mediante técnicas de data augmentation fue implementada con el fin de incrementar artificialmente el tamaño y la variabilidad del dataset de imágenes del Test de la Figura Compleja de Rey (TFCR). Esta estrategia resulta fundamental en contextos clínicos donde la cantidad de datos disponibles es limitada, ya que permite mejorar la capacidad de generalización del modelo sin necesidad de recolectar nuevas muestras, lo cual suele ser costoso y logísticamente complejo (Shorten & Khoshgoftaar, 2019).

En el código desarrollado, se aplicaron transformaciones controladas como la inversión horizontal aleatoria (`transforms.RandomHorizontalFlip()`), que introduce variaciones en la disposición espacial de los trazos, manteniendo la estructura esencial de la figura y su valor diagnóstico. Adicionalmente, el redimensionamiento homogéneo (`transforms.Resize(224,224)`) y la conversión a escala de grises normalizada aseguran que estas variaciones se integren de manera coherente con el resto del dataset.

El uso de data augmentation en el entrenamiento ayuda a mitigar el riesgo de sobreajuste, ya que expone al modelo a múltiples variaciones de una misma clase, favoreciendo la robustez en la detección de patrones relevantes. Desde el punto de vista médico, estas técnicas permiten que el sistema aprenda a reconocer el deterioro cognitivo independientemente de ligeras diferencias en la orientación, tamaño o estilo del trazo, lo que refleja mejor la diversidad real de los pacientes (Perez & Wang, 2017).

La aplicación de data augmentation, garantiza que el conjunto de datos sea más diverso, equilibrado y representativo, mejorando así el desempeño del modelo en la clasificación automática del deterioro cognitivo a partir del TFCR y aumentando la fiabilidad de la herramienta en un entorno clínico real.

- División del dataset en subconjuntos de entrenamiento, validación y prueba.

La división del dataset en subconjuntos de entrenamiento, validación y prueba constituye un paso esencial para garantizar la correcta evaluación del modelo y evitar problemas de sobreajuste. En este proyecto, el dataset depurado de 1172 imágenes fue organizado en tres particiones: entrenamiento, utilizado para ajustar los parámetros del modelo; validación, empleado para supervisar el rendimiento durante el entrenamiento y ajustar hiperparámetros; y prueba, reservado exclusivamente para evaluar el desempeño final del sistema en datos nunca vistos.

En el código implementado, la división se operacionalizó a través de la librería torchvision.datasets.ImageFolder, organizando las imágenes en carpetas diferenciadas para Train y Test. El conjunto de validación se derivó de los datos de entrenamiento mediante un muestreo controlado, lo que permitió ajustar la arquitectura sin exponer al modelo a los datos de prueba. Además, se utilizó el DataLoader con shuffle=True en la fase de entrenamiento para garantizar que las muestras se presentaran en orden aleatorio, reduciendo la posibilidad de sesgos por secuencias fijas (Goodfellow, Bengio, & Courville, 2016).

Desde la perspectiva metodológica, esta estrategia se sustenta en que la evaluación sobre un conjunto independiente de prueba es la única forma de obtener una estimación imparcial del rendimiento del modelo en situaciones reales (Kohavi, 1995). Asimismo, la inclusión de un conjunto de validación permite

controlar el sobreajuste (overfitting), es decir, el aprendizaje excesivo de los patrones específicos de entrenamiento que no generalizan adecuadamente a nuevos datos (Srivastava et al., 2014).

De esta manera, la división del dataset asegura que el modelo desarrollado para la detección temprana del deterioro cognitivo mantenga un buen equilibrio entre ajuste y generalización, garantizando su aplicabilidad en entornos clínicos reales.

## **Fase 2: Diseño de la red neuronal e implementación de arquitecturas XAI**

Objetivo asociado: Diseñar e implementar técnicas de aprendizaje profundo explicable.

Actividades:

- Selección de arquitecturas base para clasificación de imágenes: CNNs, Vision Transformers o modelos híbridos.

En el marco de la selección de arquitecturas base para la clasificación de imágenes, se optó inicialmente por el uso de redes neuronales convolucionales (CNNs), en particular la arquitectura ResNet-18 pre entrenada. La decisión responde tanto a consideraciones técnicas como al contexto del proyecto.

Las CNNs han demostrado ser altamente efectivas en tareas de clasificación de imágenes debido a su capacidad para extraer jerárquicamente características locales y globales mediante convoluciones. En el caso del Test de la Figura Compleja de Rey (TFCR), donde el trazo, la forma y la disposición espacial son determinantes para identificar deterioro cognitivo, esta característica resulta esencial. Los filtros

convolucionales permiten detectar bordes, contornos y patrones geométricos, elementos clave en la evaluación neuropsicológica automatizada (LeCun, Bengio, & Hinton, 2015; Gonzalez & Woods, 2018).

Desde el punto de vista práctico, el código implementado utiliza transfer learning con ResNet-18, una CNN entrenada previamente en el conjunto masivo de imágenes de ImageNet. Esto ofrece dos ventajas inmediatas: en primer lugar, el modelo ya dispone de representaciones visuales generales que se adaptan bien a nuevos dominios con menos datos (como en nuestro caso, con 1172 imágenes filtradas); y en segundo lugar, permite reducir el tiempo y los recursos computacionales de entrenamiento, al congelar capas convolucionales ya optimizadas y ajustar únicamente la capa final de clasificación (`model.fc = nn.Linear(num_features, 3)` en el código).

Si bien alternativas más recientes como los Vision Transformers (ViTs) han mostrado resultados sobresalientes en grandes volúmenes de datos, su rendimiento suele ser menos robusto en datasets reducidos debido a la ausencia de inductive bias espacial inherente a las CNNs, lo cual exige una mayor cantidad de muestras y recursos computacionales para generalizar adecuadamente (Dosovitskiy et al., 2020). De manera similar, los modelos híbridos que combinan convoluciones y mecanismos de atención ofrecen un potencial prometedor, pero requieren experimentación más extensa, lo que supera el alcance inicial de este proyecto.

Por estas razones, comenzar con una CNN como ResNet-18 constituye una decisión metodológica equilibrada, que permite obtener resultados sólidos y clínicamente interpretables, aprovechando un balance óptimo entre precisión, eficiencia computacional y aplicabilidad en entornos de salud. Esto sienta

la base para, en fases futuras, comparar el desempeño con arquitecturas más complejas como ViTs o híbridos, en línea con el objetivo de explorar diferentes alternativas de detección automática del deterioro cognitivo.

Tabla 2. Tabla comparativa de modelos de análisis de imagen.

<b>Criterio</b>	<b>CNN (ResNet, etc.)</b>	<b>Vision Transformers (ViTs)</b>	<b>Modelos Híbridos (CNN + ViTs)</b>
<b>Inductive bias (sesgo inductivo)</b>	Integran de manera natural la invariancia espacial (detectan bordes, formas, texturas). Ideal para trazos y patrones en imágenes del TFCR.	No poseen sesgo espacial incorporado → requieren más datos para aprender relaciones entre píxeles.	Combinan lo mejor de ambos mundos: convoluciones para detalles locales y atención para relaciones globales.
<b>Requerimiento de datos</b>	Funcionan bien con datasets pequeños o medianos gracias al transfer learning.	Requieren grandes volúmenes de datos para superar a CNNs (Dosovitskiy et al., 2020).	Menor necesidad de datos que ViTs puros, pero mayor que CNNs tradicionales.

<b>Costo computacional</b>	Bajo-medio. ResNet-18 puede entrenarse en GPU estándar con eficiencia.	Alto. ViTs demandan más GPU, memoria y tiempo de entrenamiento.	Medio-alto. Más costosos que CNNs, pero menos que ViTs puros.
<b>Interpretabilidad clínica</b>	Compatible con técnicas de XAI como Grad-CAM o LIME, que destacan trazos relevantes del dibujo (Gunning & Aha, 2019).	Más difícil de interpretar debido a la complejidad de los mecanismos de atención.	Mejoran la interpretabilidad al combinar mapas de características (CNN) con atención visual (ViT).
<b>Criterio</b>	<b>CNN (ResNet, etc.)</b>	<b>Vision Transformers (ViTs)</b>	<b>Modelos Híbridos (CNN + ViTs)</b>
<b>Generalización en contextos clínicos</b>	Robusta incluso con datos limitados; probada en aplicaciones médicas (Litjens et al., 2017).	Prometedora en grandes datasets médicos, pero aún experimental en dominios reducidos.	Potencial alto en el futuro aunque todavía en experimental.
<b>Justificación para el proyecto</b>	Balance óptimo entre precisión, eficiencia y aplicabilidad clínica en dataset reducido (~1172 imágenes).	No es la mejor opción inicial por el tamaño limitado de datos y recursos disponibles.	Interesante para futuras fases de comparación, pero prioritario en la actualidad.

## Proceso de Entrenamiento del Modelo

El entrenamiento del modelo se llevó a cabo mediante una estrategia de transfer learning empleando la arquitectura ResNet-18 preentrenada en ImageNet, adaptada para clasificar las tres categorías definidas en el dataset del Test de la Figura Compleja de Rey (TFCR).

Tabla 3. Comparación de redes neuronales

<b>Criterio</b>	<b>Peso</b>	<b>Puntaje ResNet-18 (1-5)</b>	<b>Puntaje ponderado ResNet-18</b>	<b>Puntaje EfficientNet-B0 (1-5)</b>	<b>Puntaje ponderado EfficientNet-B0</b>	<b>Puntaje Vision Transformer (ViT) (1-5)</b>	<b>Puntaje ponderado Vision Transformer (ViT)</b>
Desempeño con datos limitados (transfer learning)	0.18	5	0.9	4	0.72	3	0.54

Capacidad para rasgos de contorno/forma en dibujos de línea	0.16	5	0.8	4	0.64	3	0.48
Facilidad de fine-tuning selectivo (layer3/layer4)	0.16	5	0.8	4	0.64	3	0.48

Compatibilidad y madurez con Grad-CAM	0.16	5	0.8	4	0.64	3	0.48
---------------------------------------	------	---	-----	---	------	---	------

Eficiencia computacional (entrenamiento en CPU/recursos moderados)	0.14	5	0.7	4	0.56	2	0.28
Estabilidad y reproducibilidad del entrenamiento	0.1	5	0.5	4	0.4	3	0.3
Disponibilidad de implementaciones y pesos preentrenados	0.1	5	0.5	5	0.5	4	0.4

TOTAL	1		5		4.1		2.96
-------	---	--	---	--	-----	--	------

Se eligió ResNet-18 por su solidez en transfer learning con pocos datos, su compatibilidad directa con Grad-CAM y su eficiencia para entrenar/ajustar selectivamente capas profundas en el contexto de dibujos de línea con alta similitud.

En primer lugar, las capas convolucionales de ResNet-18 fueron congeladas (`param.requires_grad = False`), de modo que los pesos previamente aprendidos en tareas de reconocimiento general de imágenes permanecieran fijos. Esto permitió aprovechar representaciones visuales robustas (bordes, formas, patrones geométricos), reduciendo tanto el tiempo de entrenamiento como la necesidad de un dataset masivo (LeCun, Bengio, & Hinton, 2015). Únicamente se reentrenó la capa totalmente conectada final (`model.fc = nn.Linear(num_features, 3)`), adaptándola al problema específico de clasificación en tres clases: normal, deterioro cognitivo leve y deterioro cognitivo avanzado. El proceso de entrenamiento se estructuró en épocas (`epochs=15`), donde en cada una el modelo pasó por dos fases:

Fase de entrenamiento:

- El modelo se colocó en modo entrenamiento (`model.train()`), procesando lotes de imágenes de 32 muestras (`batch_size=32`).
- Cada lote pasó por un ciclo de forward pass, cálculo de pérdida con entropía cruzada (`criterion = nn.CrossEntropyLoss()`), retropropagación del error (`backpropagation`) y actualización de los parámetros de la capa final mediante el optimizador Adam (`optim.Adam`) con una tasa de aprendizaje de 0.001.

- Durante esta fase se registraron métricas de desempeño como la pérdida promedio y la precisión sobre los datos de entrenamiento.

Fase de validación:

- El modelo se colocó en modo evaluación (`model.eval()`), deshabilitando la actualización de pesos.
  - Se procesaron los datos del conjunto de prueba, calculando la precisión global sin modificar los parámetros aprendidos.
- Esta fase permitió verificar la capacidad del modelo de generalizar a datos no vistos y detectar posibles problemas de sobreajuste (Srivastava et al., 2014).

A lo largo del proceso, se recolectaron métricas de pérdida, precisión en entrenamiento y precisión en prueba, las cuales se graficaron para analizar la evolución del aprendizaje. Esta práctica permite identificar la convergencia del modelo y evaluar el balance entre ajuste y generalización (Goodfellow, Bengio, & Courville, 2016).

Finalmente, el modelo entrenado fue guardado en formato `.pth` (`torch.save(model.state_dict(), 'modelo_dibujos.pth')`), permitiendo su reutilización en predicciones posteriores sin necesidad de repetir todo el entrenamiento.

Para la decisión sobre la herramienta para la construcción de la red neuronal convolucional, se evaluaron 3 posibles candidatos, los cuales fueron PyTorch, TensorFlow/Keras y finalmente `fgastai`, para lo cual se decidió mediante la siguiente matriz el uso de PyTorch.

Tabla 4. Comparación de librerías para preparación de la red neuronal

Criterio	Peso	Puntaje PyTorch (1-5)	Puntaje ponderado PyTorch	Puntaje TensorFlow w/Keras (1-5)	Puntaje ponderado TensorFlow/Keras	Puntaje fastai (1-5)	Puntaje ponderado fastai
Integración con el pipeline actual (código, DataLoader, hooks)	0.18	5	0.9	4	0.72	4	0.72
Flexibilidad para investigación/experimentos	0.16	5	0.8	4	0.64	4	0.64
Soporte para AI (hooks, gradients, capas)	0.16	5	0.8	4	0.64	4	0.64
Comunidad, documentación y ecosistema	0.14	5	0.7	5	0.7	4	0.56
Eficiencia en CPU/GPU y control fino del entrenamiento	0.14	5	0.7	4	0.56	4	0.56

Reproducibilidad (control de seeds, loaders, determinismo)	0.	5	0.6	4	0.48	4	0.48
Compatibilidad con bibliotecas médicas/visuales	0.	5	0.5	4	0.4	4	0.4
TOTAL	1		5		4.14		4

Se eligió PyTorch por su integración directa con el código existente, su control granular del entrenamiento y su soporte nativo para técnicas XAI basadas en gradientes y activaciones, manteniendo reproducibilidad y eficiencia en CPU/GPU.

El primer entrenamiento del modelo se realizó utilizando un dataset sintético conformado por imágenes de los números 1, 2 y 3 en diferentes tipografías, con el propósito de validar la arquitectura seleccionada (ResNet-18 con transfer learning) y comprobar el correcto funcionamiento del pipeline de preprocesamiento antes de trabajar con los dibujos del Test de la Figura Compleja de Rey (TFCR). Esta etapa inicial permitió evaluar la capacidad del modelo para reconocer patrones estructurales básicos bajo condiciones controladas.

En la gráfica de la izquierda, correspondiente a la pérdida durante entrenamiento, se observa una disminución progresiva desde valores cercanos a 1.5 hasta alcanzar aproximadamente 0.7 al final de las

épocas. Esta tendencia confirma que el modelo va reduciendo de manera estable los errores de predicción a medida que ajusta los parámetros de la capa final. El hecho de que la curva sea descendente y sin oscilaciones abruptas refleja que el uso del optimizador Adam con una tasa de aprendizaje de 0.001 permitió una convergencia adecuada.

La gráfica de la derecha muestra la precisión en entrenamiento y prueba. Inicialmente, la precisión en entrenamiento es baja (~30%), debido a que la capa final parte de pesos aleatorios y requiere varias iteraciones para ajustarse a las variaciones tipográficas. Sin embargo, a partir de la mitad del entrenamiento se evidencia un incremento sostenido, alcanzando más del 70% de precisión en ambas particiones. Un aspecto relevante es que la precisión del conjunto de prueba se mantuvo consistentemente ligeramente superior a la de entrenamiento. Esto se explica porque el conjunto de entrenamiento fue sometido a técnicas de data augmentation, como la inversión horizontal aleatoria, que introducen mayor variabilidad y complejidad; mientras que el conjunto de prueba, al no estar aumentado, resultó más sencillo de clasificar. Este comportamiento evidencia que el modelo generalizó de manera adecuada sin mostrar indicios de sobreajuste.

En términos metodológicos, este primer experimento cumplió la función de validación preliminar del código y de la estrategia de transfer learning. El buen desempeño obtenido con un dataset sencillo, pero con variaciones tipográficas, demuestra que la red neuronal convolucional es capaz de abstraer los rasgos estructurales esenciales de las imágenes y clasificarlos con un nivel de precisión aceptable. En consecuencia, este resultado respalda la pertinencia de avanzar hacia el análisis de los dibujos del TFCR, donde las diferencias entre clases son más sutiles y clínicamente significativas.

Este procedimiento metodológico se encuentra respaldado por la literatura en visión por computador, donde es común el uso de datasets sintéticos o de menor complejidad para probar arquitecturas antes de abordar datos más complejos. Según Tremblay et al. (2018), los datos sintéticos permiten validar la robustez de modelos y reducir riesgos iniciales de sobreajuste, mientras que Qi et al. (2016) destacan que estos experimentos piloto facilitan la detección temprana de fallos en preprocesamiento, optimización y arquitectura. De igual forma, Goodfellow, Bengio y Courville (2016) subrayan que iniciar con tareas controladas favorece la transferencia progresiva del aprendizaje hacia escenarios más desafiantes.



*Ilustración 6. Números utilizados para el entrenamiento. (Fuente. Google imágenes.2025)*

- Ajuste de hiper parámetros (optimización, número de capas, tasa de aprendizaje).

Luego de las pruebas del modelo de comparación y análisis de imágenes entrenado mediante diferentes imágenes de números seleccionando entre 3 números posibles (1, 2 y 3), se cargó el modelo inicial de red neuronal con el dataset de las 1172 imágenes para evaluar su desempeño y de esta manera analiza el desempeño en fase de entrenamiento y prueba.

Para ello se separaron los grupos de imágenes correspondientes a 3 grupos de clasificación realizada por profesionales para los cuales se encuentran divididos en grupos de 0, 1 y 2 que corresponden a:

- 0 = Normal
- 1 = Deterioro cognitivo leve
- 2 = Demencia

En esta fase se implementó el mismo modelo de red neuronal utilizado con los números, buscando un entrenamiento y pruebas capaces de mostrar el avance con un modelo básico de procesamiento neuronal mediante las herramientas de PYTorch.

Para la fase de entrenamiento con la base de datos se utilizó una repartición de recursos del dataset de 90%/5%/5% (Entrenamiento, Validación, Pruebas), con la finalidad de tener un modelo entrenado cuyo resultado ayudará a la orientación en cómo mejorar y posterior a ello agregar la inteligencia artificial explicativa (XAI), en busca de mejora para el modelo afinando los puntos clave en los que se quiere enfocar el análisis para la clasificación mediante la red neuronal.

En este modelo se utilizó una red neuronal con 10 épocas con la finalidad de minimizar la pérdida en el entrenamiento. La cantidad de épocas (pasadas completas del conjunto de entrenamiento a través de la red neuronal) afecta directamente el rendimiento en tareas de comparación de imágenes (como reconocimiento de similitudes, búsqueda de duplicados o verificación de identidad). El modelo que se entrenó, se guardó para ser analizado posteriormente en el archivo 'modelo\_dibujos.pth'.

El resultado del entrenamiento a 10 épocas fue el siguiente:

Época [1/10]

Pérdida: 1.0163, Precisión Train: 54.73%, Precisión Test: 44.32%

---

Época [2/10]

Pérdida: 0.9125, Precisión Train: 58.90%, Precisión Test: 45.45%

---

Época [3/10]

Pérdida: 0.8779, Precisión Train: 61.27%, Precisión Test: 43.18%

---

Época [4/10]

Pérdida: 0.8648, Precisión Train: 60.98%, Precisión Test: 45.45%

---

Época [5/10]

Pérdida: 0.8325, Precisión Train: 62.69%, Precisión Test: 45.45%

---

Época [6/10]

Pérdida: 0.8318, Precisión Train: 62.31%, Precisión Test: 44.32%

---

Época [7/10]

Pérdida: 0.8324, Precisión Train: 64.39%, Precisión Test: 55.45%

---

Época [8/10]

Pérdida: 0.8380, Precisión Train: 62.97%, Precisión Test: 40.91%

---

Época [9/10]

Pérdida: 0.8052, Precisión Train: 64.39%, Precisión Test: 40.91%

---

Época [10/10]

Pérdida: 0.8058, Precisión Train: 66.00%, Precisión Test: 44.32%

---

Tabla 5. Resumen de entrenamiento por épocas utilizando el modelo inicial. (Autoría propia)

Época	Pérdida	Train Acc	Test Acc	Observación
1	10.163	54.73%	44.32%	Comportamiento esperado inicial
		↑	↓	
2	0.9125	58.90%	45.45%	Mejor test (punto óptimo potencial)
		↑	↓	
3	0.8779	61.27%	↓ 43.18%	Primera señal de sobreajuste
		↑	↓ ↔	
5	0.8325	62.69%	55.45%	Estancamiento de test
		↓	↑	
8	0.8380	62.97%	↓ 40.91%	Sobreajuste severo
	↔	↑		
10	0.8058	66.00%	↑ 44.32%	Recuperación insuficiente

Durante el entrenamiento de la red neuronal para comparación de imágenes, se observó un fenómeno de sobreajuste (*overfitting*) significativo. Los registros de las 10 épocas muestran que mientras la precisión en entrenamiento aumentó consistentemente del 54.73% al 66.00%, la precisión en el conjunto de validación se estancó en un rango de 40.91%–55.45%, sin superar el máximo de 55.45% alcanzado en las épocas 2 y 5. Esta divergencia progresiva (23.7% en época 10) indica que el modelo optimizó su ajuste a los datos de entrenamiento a expensas de su capacidad de generalización (Prechelt, 1998).

Se identificó un punto óptimo en la época 5 (pérdida: 0.8325; precisión validación: 55.45%), tras el cual el rendimiento en validación decayó hasta 40.91% en épocas posteriores, evidenciando una degradación por sobre entrenamiento. Este patrón sugiere que el modelo desarrolló alta sensibilidad a ruido y variaciones irrelevantes en los datos de entrenamiento (Hawkins, 2004), lo que se manifiesta en:

- Alta varianza en precisión de validación ( $\pm 4.54\%$  entre épocas)

- Desconexión entre la reducción de pérdida en entrenamiento y el rendimiento en validación

Dado lo anterior, se corrió un código que carga el modelo entrenado y genera una predicción de acuerdo con una imagen que se analiza a fin de evaluar la veracidad del modelo, en donde los resultados fueron los siguientes:

Predicción: Clase2

<b>Imagen_id</b>	<b>Edad</b>	<b>Escolaridad</b>	<b>Sexo</b>	<b>DX</b>
1172	83	4	2	2

Caracterización de datos para la imagen 1172 analizada en la predicción de clase.

Predicción: Clase2

<b>Imagen_id</b>	<b>Edad</b>	<b>Escolaridad</b>	<b>Sexo</b>	<b>DX</b>
1167	76	2	1	1

Caracterización de datos para la imagen 1167 analizada en la predicción de clase. (Autoría propia).

Predicción: Clase2

<b>Imagen_id</b>	<b>Edad</b>	<b>Escolaridad</b>	<b>Sexo</b>	<b>DX</b>
1168	52	4	2	0

Caracterización de datos para la imagen 1167 analizada en la predicción de clase. (Autoría propia).

Ante el resultado de solo 1 acierto en 3 imágenes evaluadas (33.3% de precisión), se evidencia que el modelo actual presenta limitaciones críticas de generalización y robustez. Este desempeño, consistente con los patrones de sobreajuste observados durante el entrenamiento, obliga a implementar mejoras sistémicas.

Teniendo en cuenta que la finalidad es implementar herramientas de XAI en pro de buscar una mejora para afinar los modelos de las redes neuronales propuestas de tal forma que se pueda llegar a obtener una herramienta estable y utilizable, se planteo el mejorar el modelo de entrenamiento previo a la implementación de XAI en busca de mejora en el rendimiento.

Para ello se tuvo en cuenta lo siguiente:

- Incrementar la cantidad de épocas para mejorar el modelo.
- Se debe incrementar en la medida de lo posible la cantidad de imágenes que se tiene en el dataset para las categorías Normal y Demencia. Esto se debe a que la repartición de recursos de las imágenes se compone de la siguiente forma: o Normal: 214 imágenes o Deterioro cognitivo leve: 665 imágenes o Demencia: 287 imágenes
- Se debe enseñar al modelo a reconocer objetos en condiciones reales variables.
- Utilizar un modelo de Fine Tuning mejorado, ya que las capas iniciales aprenden bordes/texturas, las finales conceptos específicos.
- Se debe mejorar la gestión de entrenamiento, se hizo uso de métodos como:
  - o Learning Rate Warmup para las primeras 5 épocas
  - o Cosine Annealing (20-30 épocas) o Early Stopping con monitoreo. Esto con la finalidad de disminuir la cantidad de recursos usados en caso de que el modelo no presente mejoras y supere la tolerancia a utilizar.
- Se añade una matriz de confusión para los resultados en el próximo modelo

- Se añade el uso de XAI para la visualización del funcionamiento de la red neuronal.

Luego de aplicar correcciones en el código y generar un nuevo modelo de entrenamiento, aplicando las consideraciones mencionadas, se obtuvieron los siguientes resultados dando un stop por tolerancia de 7 puntos en 8 épocas: Dispositivo: cpu

Clases: ['0', '1', '2']

[Época 1/25] Loss: 0.9836 | Acc Train: 55.98% | Acc Val: 64.45%

↳ Nuevo mejor modelo guardado: outputs\best\_model.pth

[Época 2/25] Loss: 0.7685 | Acc Train: 64.02% | Acc Val: 60.66%

[Época 3/25] Loss: 0.5617 | Acc Train: 69.70% | Acc Val: 62.56%

[Época 4/25] Loss: 0.4554 | Acc Train: 74.08% | Acc Val: 60.19%

[Época 5/25] Loss: 0.3351 | Acc Train: 79.88% | Acc Val: 54.98%

[Época 6/25] Loss: 0.4651 | Acc Train: 74.67% | Acc Val: 61.61%

[Época 7/25] Loss: 0.3729 | Acc Train: 71.95% | Acc Val: 59.24% [Época

8/25] Loss: 0.2449 | Acc Train: 79.88% | Acc Val: 62.56%

Early stopping activado (paciencia=7)

precision recall f1-score support

0 0.0000 0.0000 0.0000 24

1 0.4638 0.8000 0.5872 40

2 0.3158 0.2500 0.2791 24

accuracy 0.4318 88

macro avg 0.2599 0.3500 0.2887 88 weighted

avg 0.2969 0.4318 0.3430 88

El proceso se ejecutó en CPU, utilizando un clasificador basado en ResNet para tres clases (['0', '1', '2']).

El entrenamiento se configuró para 25 épocas, con criterio de early stopping activado tras detectar ausencia de mejora en la métrica de validación durante siete iteraciones consecutivas.

## **Integración de técnicas de inteligencia artificial explicable (XAI) como Grad-CAM, LIME o SHAP para generar visualizaciones interpretables.**

La integración de técnicas de Inteligencia Artificial Explicable (XAI) en entornos clínicos responde a la necesidad crítica de garantizar transparencia, trazabilidad y confianza en sistemas basados en aprendizaje profundo. En el ámbito de la salud, donde las decisiones automatizadas impactan directamente en la vida de los pacientes, la opacidad de los modelos de caja negra plantea riesgos éticos y regulatorios significativos. XAI permite comprender cómo los algoritmos procesan la información y qué factores influyen en sus predicciones, favoreciendo la validación clínica y la adopción responsable de estas tecnologías. Diversos estudios han demostrado que la explicabilidad no solo incrementa la confianza del profesional, sino que también actúa como mecanismo para detectar sesgos y mejorar la robustez del modelo (Frasca et al., 2024; Nzenwata et al., 2024). En este contexto, técnicas como Grad-CAM y LIME se consolidan como herramientas esenciales para interpretar modelos de visión artificial, aportando explicaciones visuales y locales que facilitan la comprensión del proceso de decisión (Selvaraju et al., 2017; Ribeiro et al., 2016).

*Tabla 6. Comparativa de XAI disponibles*

<b>Técnica</b>	<b>Modelo</b>	<b>Tipo</b>	<b>Dominio ideal</b>	<b>Fidelidad al modelo</b>	<b>Granularidad</b>	<b>Costo computacional</b>	<b>Ventajas principales</b>	<b>Limitaciones</b>	<b>Razón/uso recomendado</b>

<b>Grad-CAM</b>	No	Basada en gradientes (modelo-específica para CNN)	Imágenes médicas con CNN (clasificación)	Alta	Local (mapa de calor por clase)	Bajo-Medio	Mapas clase-discriminativos sin reentrenamiento; buena alineación con estructuras anatómicas; fácil superposición	No aplicable a modelos no convolucionales; depende de la última capa conv; explica señales positivas	Validar que la red atiende al dibujo/estructura; auditoría rápida en TFCR
<b>LIME</b>	Sí	Perturbación local + modelo sustituto	Cualquier dominio (imágenes/tabular/texto); útil en	Mediana	Local (superpíxeles/segmentos)	Medio-Alto	Independiente del modelo; explicación por segmentación	Sensibilidad a la segmentación/semillas; coste mayor; puede ser	Completar Grad-CAM con verificación

			<b>inferencia por caso</b>				<b>os intuitiva; útil para validar coherencia</b>	<b>inestable entre corridas</b>	<b>agnóstica por caso</b>
SHA P	Sí	Teoría de juegos (atribuciones)	Tabular; también imágenes con wrappers	Alta (consistencia teórica)	Local y global (agregable)	Alto (especialmente Kernel/DeepS HAP en alta dimensión)	Atribuciones consistentes; permite rankings globales de características; trazabilidad	Coste elevado; requiere supuestos de independencia; para imágenes puede ser pesado	Análisis de factores clínicos tabulares o mezcla imagen +clínico
Integrated Gradients	No	Basada en gradientes	Redes profundas diferenciab les (incluye CNN)	Alta	Local (atribución por píxel)	Medio	Evita problemas de saturación;	Depende de la elección de baseline;	Atribución fina cuando se necesita

		acumulados					axiomas claros; no requiere reentrenar	mapas menos intuitivos para clínicos sin postprocesado	precisión por píxel
Grad-CAM++	No	Basada en gradientes (mejora de Grad-CAM)	CNN en imágenes donde hay múltiples instancias finas	Alta	Local (mapa de calor por clase)	Medio	Mejor localización para objetos múltiples/pequeños; más estable ante variaciones	Más costosa que Grad-CAM; implementaciones menos extendidas	Cuando se requiere mayor precisión espacial que Grad-CAM
RISE	Sí	Perturbación estocástica	Imágenes (cualquier modelo)	Mediana-Alta	Local (importancia probabilística)	Alto	Agnóstica y simple; buenas	Muy costosa (múltiples evaluaciones)	Validación robusta offline

		(enmas caramie nto aleatori o)			ca por píxel)		propieda des de fidelida d en benchm arks	nes); mapas ruidosos si no se promedia suficiente	cuando el coste no es crítico
Occl usion Sensi tivity	Sí	Perturb ación determi nística (oclusi ones)	Imágenes (cualquier modelo)	Medi a	Local (búsqueda por ventanas)	Medio- Alto	Intuitiva ; control directo del tamaño de ventana	Coste alto; artefactos por tamaño/fo rma de máscara; sensible al stride	Análisis de ablación sencillo para casos específi cos
Smo othG rad	No	Suaviza do de gradien tes (prome dio con ruido)	Modelos diferenciab les (CNN/DN N)	Medi a- Alta	Local (saliency suavizada)	Medio	Reduce ruido en mapas de gradient e; fácil de aplicar	No corrige sesgos del método base; añade costo por múltiples pases	Mejorar legibilid ad de mapas de gradient e/IG

La elección de Grad-CAM y LIME se fundamentó en su complementariedad metodológica: mientras Grad-CAM ofrece una explicación fiel al modelo mediante mapas de activación que revelan las regiones más

influyentes en la predicción, LIME aporta una perspectiva agnóstica basada en perturbaciones locales, permitiendo validar la coherencia de las decisiones desde un enfoque independiente. Esta combinación fortaleció la validez interna de las interpretaciones y garantizó que las explicaciones fueran tanto precisas como comprensibles para el personal clínico, alineando el desarrollo del sistema con los principios de transparencia y trazabilidad exigidos en aplicaciones sanitarias.

Tabla 7. Justificación de selección de Grad-CAM y LIME

<b>Técnica</b>	<b>Tipo</b>	<b>Ventajas principales</b>	<b>Limitaciones</b>	<b>Razón de selección</b>
<b>Grad-CAM</b>	<b>Basada en gradientes (modelo-específica)</b>	<b>Genera mapas de calor clase-discriminativos sin reentrenamiento; alta fidelidad al modelo; útil para CNN en imágenes médicas.</b>	<b>Depende de la arquitectura; no es agnóstica al modelo.</b>	<b>Permite validar que la red atiende a regiones relevantes del dibujo, asegurando coherencia clínica.</b>
<b>LIME</b>	<b>Basada en perturbaciones (agnóstica al modelo)</b>	<b>Explica predicciones individuales mediante superpíxeles; independiente del tipo de modelo; fácil interpretación local.</b>	<b>Mayor coste computacional; sensibilidad a parámetros de segmentación.</b>	<b>Complementa Grad-CAM al ofrecer explicaciones locales y validar coherencia desde un enfoque externo al modelo.</b>

Con el propósito de hacer auditable y trazable el proceso de decisión del clasificador de imágenes, se desarrolló un módulo de Inteligencia Artificial Explicable (XAI) que integra Grad-CAM y LIME como técnicas complementarias. Grad-CAM permite localizar espacialmente las regiones de mayor contribución a la predicción de una clase en modelos convolucionales, proyectando gradientes sobre la última capa convolucional y produciendo mapas de calor clase-discriminativos sin reentrenamiento (Selvaraju et al., 2017). LIME explica predicciones individuales mediante perturbaciones locales a manera de superpíxeles y ajuste de un modelo interpretable que estima la importancia de segmentos visibles sobre la salida del clasificador (Ribeiro et al., 2016). La combinación de ambos enfoques incrementa la validez interna de la interpretación al converger evidencias desde una técnica fiel al modelo (Grad-CAM) y otra agnóstica (LIME), práctica recomendada en la literatura de XAI para equilibrar fidelidad y comprensibilidad (Ribeiro et al., 2016; Selvaraju et al., 2017).

El script XAI.py consume los artefactos de entrenamiento generados por el pipeline principal: los pesos del modelo (best\_model.pth o last\_model.pth), el mapeo de clases (classes.json) y, cuando está disponible, el parámetro de temperatura (temperature.json) para calibrar la confianza. La calibración por temperature scaling corrige la sobre confianza típica de redes modernas y alinea la probabilidad reportada con la verosimilitud real de acierto (Guo et al., 2017). En consecuencia, todas las probabilidades que se muestran en el módulo XAI pueden aplicarse sobre logits reescalados por T, mejorando la interpretación de la fiabilidad (Guo et al., 2017).

### **Fase 3: Evaluación del desempeño de los modelos**

Objetivo asociado: Evaluar el desempeño de las arquitecturas desarrolladas.

Actividades:

- Medición de métricas de clasificación: accuracy, precisión, recall, F1-score.

La combinación accuracy + matriz de confusión + reporte de clasificación permite evaluar performance global (accuracy), por clase (P/R/F1) y patrones de error (confusión entre etiquetas). Además, el script incorpora calibración de probabilidades (temperature scaling), y reevalúa con T óptima para obtener métricas bajo probabilidades más fieles (aunque la calibración, por sí misma, no altera los conteos de aciertos, sí afecta métricas dependientes de umbrales si se usaran).

- Accuracy (global): en la función `evaluate_and_report(...)`, tras consolidar `all_preds` y `all_labels`, se calcula la exactitud como el promedio de aciertos (`(all_preds == all_labels).mean() * 100`) y se reporta en consola, incluyendo la temperatura usada para la evaluación (calibrada o no)
  - Matriz de confusión: en la misma función se construye `cm` (conteos por verdadero vs. predicho), se gráfica y guarda en `outputs/confusion_matrix_<split>.png`. Esta matriz respalda análisis por clase y detección de confusiones sistemáticas.
  - Precisión, recall y F1 por clase: el código invoca `sklearn.metrics.classification_report(...)` para generar el reporte completo (precisión, recall, F1-score por clase, macro y weighted), lo imprime y además lo guarda en `outputs/classification_report_<split>.txt`. Con ello se cubren las métricas solicitadas a nivel por-clase y promedios macro/ponderado.
- Evaluación de eficiencia computacional (inferencia, uso de memoria).
  - Entrenamiento: tiempo total, tiempo medio por época, ejemplos/segundo.
  - Inferencia: ms/imagen en test y/o en el flujo de inferencia libre (INFER\_ENTRY).
  - Memoria: pico de memoria reservada/allocada (GPU) y RSS (RAM).

- Análisis de capacidad de generalización a partir del rendimiento en el conjunto de prueba.

Tras el entrenamiento y selección del mejor modelo (según Acc Val), el script ejecuta `evaluate_and_report(...)` en el conjunto de prueba y reporta:

- Accuracy global en test (con y sin calibración por temperatura),
- Matriz de confusión y reporte de clasificación (precisión/recall/F1 por clase),
- ECE (Expected Calibration Error) para evaluar la calibración de confianza.
- Además, cuando `EXPORT_TEST_CSV=True`, genera un CSV por imagen con etiqueta verdadera, predicción, probabilidad de la clase predicha, Top-K y la temperatura empleada, lo que facilita auditorías y análisis de errores difíciles fuera de consola.

Generalización implica medir el desempeño en un conjunto no visto durante el entrenamiento. Aquí, el pipeline compara Acc Val (durante training) con Acc Test y analiza distribuciones por clase vía la matriz de confusión y el reporte por clase (P/R/F1). Diferencias notables entre validación y prueba pueden evidenciar shift de distribución, sobreajuste o escasez/desbalance de datos en ciertas clases.

La métrica ECE aporta la dimensión de confiabilidad: incluso si la exactitud es estable, una ECE alta sugiere que el porcentaje de confianza mostrado por el modelo no refleja su probabilidad real de acierto, útil para determinar umbrales de rechazo y criterios de revisión humana.

En la fase de evaluación del desempeño se implementó un conjunto de procedimientos orientados a medir la capacidad predictiva y la generalización de las arquitecturas desarrolladas. El análisis se centró en métricas clave como exactitud global, precisión, recall y F1-score, complementadas con matrices de confusión y reportes por clase que permitieron identificar patrones de error y confusiones sistemáticas. Además, se incorporó la calibración de probabilidades mediante escalado de temperatura para ajustar la confianza del modelo a su verosimilitud real, lo que resultó útil para definir umbrales de decisión y derivación a revisión humana en casos ambiguos. El pipeline también evaluó la eficiencia computacional, registrando tiempos de entrenamiento, inferencia y consumo de memoria, y comparó el rendimiento en validación frente al conjunto de prueba para detectar posibles fenómenos de sobreajuste o desplazamiento de distribución.

Como parte del control del proceso de entrenamiento, se utilizó un early stopper con tolerancia de 8 épocas. Esta decisión metodológica respondió a la necesidad de evitar el sobreajuste, fenómeno observado en corridas preliminares donde el modelo continuaba reduciendo la pérdida en entrenamiento sin mejorar la precisión en validación. El criterio de paciencia fijado en ocho iteraciones permitió otorgar un margen suficiente para que el optimizador explorara mejoras reales antes de detener el entrenamiento, evitando cortes prematuros que pudieran comprometer la convergencia. En la práctica, este mecanismo redujo el riesgo de que el modelo aprendiera patrones irrelevantes o ruido propio del conjunto de entrenamiento, favoreciendo una representación más generalizable. Además, el early stopping contribuyó a optimizar el uso de recursos computacionales, disminuyendo el tiempo total de ejecución y el consumo energético sin sacrificar desempeño. En síntesis, la incorporación de este control no solo mejoró la estabilidad del modelo, sino que garantizó que la selección del mejor punto de entrenamiento se basara en evidencia objetiva de rendimiento en validación, alineando el proceso con los principios de reproducibilidad y eficiencia que guiaron toda la metodología.

Tabla 8. Tabla de comparación para la selección del early stopper

Paciencia (épocas)	Ventajas esperadas	Riesgos / Costes	Observación empírica en el proyecto	Idoneidad
3	Reacciona muy rápido ante estancamiento; reduce tiempo de entrenamiento.	Alto riesgo de corte prematuro ante oscilaciones normales de validación; no deja completar ciclos del scheduler.	Se observaron mesetas cortas (<5 épocas) con rebotes posteriores; con 3 se detendría antes de recuperar.	Baja
5	Compromiso entre tiempo y exploración; protege frente a ruido leve.	Puede seguir siendo corto si la mejora aparece tras una ventana de 6–7 épocas (picos por aumentación fuerte).	En Fases 2–3 hubo oscilaciones de validación de hasta 6–7 épocas antes de un nuevo máximo.	Media
8	<b>Tolerancia suficiente para cubrir las oscilaciones típicas (5–7 épocas) y permitir que el LR scheduler y la regularización surtan efecto.</b>	<b>Ligero aumento del tiempo de entrenamiento si no hay mejora real.</b>	<b>Con 8 se evitó el sobrecorte y se capturó el mejor punto de validación sin sobreentrenar; activación observada tras no mejorar por varias épocas.</b>	<b>Alta (seleccionada)</b>
10	Mayor probabilidad de no cortar mejoras tardías; explora más el espacio de parámetros.	Riesgo de sobreajuste si la validación ya no mejora; coste computacional mayor sin ganancias claras.	No se evidenció beneficio consistente frente a 8; incremento de tiempo sin mejoras en test.	Media
12	Máxima tolerancia a mesetas largas; minimiza cortes prematuros en problemas altamente no convexos.	Aumenta notablemente el coste y la probabilidad de sobreajuste en datasets medianos; menor eficiencia operacional.	Para este dominio, no justificó el coste adicional; riesgo de memorizar ruido de entrenamiento.	Baja

La selección de una paciencia de 8 épocas respondió a un criterio metodológico orientado a equilibrar generalización y eficiencia. En corridas previas se observaron oscilaciones de la métrica de validación que se prolongaron durante ventanas de 5 a 7 épocas antes de alcanzar un nuevo máximo; por ello, fijar una paciencia de 3 o 5 habría incrementado el riesgo de corte prematuro, particularmente bajo aumentación moderada y ajustes finos de capas profundas. Por el contrario, tolerancias superiores (10–12) habrían elevado el coste computacional y la probabilidad de sobreajuste sin aportar mejoras consistentes en el conjunto de prueba, dado el tamaño y la naturaleza del dataset. En consecuencia, una paciencia de 8 permitió absorber la variabilidad esperable de la validación, sin prolongar innecesariamente el entrenamiento cuando no existían señales objetivas de mejora, y se alineó con el comportamiento del programador de tasa de aprendizaje y las regularizaciones aplicadas. En síntesis, esta decisión preservó el mejor punto de validación evitando interrupciones tempranas y, a la vez, contuvo el tiempo total de entrenamiento, reforzando la reproducibilidad y la estabilidad del modelo en la Fase 3.

#### **Fase 4: Evaluación de la interpretabilidad clínica**

Objetivo asociado: Analizar la interpretabilidad clínica de los modelos más precisos.

Actividades:

- Generación de mapas de activación y explicaciones visuales del modelo para distintos casos de prueba.

Se estableció un entorno de desarrollo basado en PyTorch y se configuraron los módulos necesarios para entrenamiento, evaluación y generación de explicaciones visuales. Se fijaron semillas aleatorias para garantizar reproducibilidad y se definieron rutas para datasets, salidas y artefactos del modelo. El dispositivo de cómputo se seleccionó dinámicamente (CPU/GPU) para optimizar el rendimiento.

Para comprender la toma de decisiones del modelo, se integraron dos técnicas de Explainable AI:

- Grad-CAM: Se implementó sobre la última capa convolucional del backbone para generar mapas de activación que indicaron las regiones más relevantes en la predicción. Estos mapas se superpusieron sobre las imágenes originales, permitiendo identificar si el modelo se centraba en el objeto o en el fondo.
- LIME (Local Interpretable Model-agnostic Explanations): Se aplicó para explicar predicciones individuales mediante perturbaciones locales y análisis de superpíxeles, revelando qué segmentos de la imagen influían más en la decisión del modelo.

Se diseñó XAI.py para operar después del entrenamiento, cargando los artefactos producidos por el pipeline (pesos del mejor modelo, lista de clases y temperatura de calibración) con el fin de generar explicaciones locales por imagen mediante Grad-CAM y LIME, y persistirlas en carpetas separadas para su análisis sistemático. Esta decisión aseguró consistencia entre el modelo utilizado en inferencia/explicación y el efectivamente validado durante entrenamiento. El script estableció rutas a `best_model.pth`, `classes.json` y `temperature.json`, y definió un punto de entrada (ENTRY) para procesar un archivo o carpeta completa.

El razonamiento del cálculo fue el siguiente: se hookearon activaciones y gradientes del módulo objetivo; ante una imagen preprocesada, se obtuvo el logit de la clase de interés y se propagó gradiente hacia atrás; se calculó un peso por canal mediante promedio global de gradientes y se combinó linealmente con las activaciones por canal; finalmente, se aplicó ReLU y se normalizó el mapa a  $[0,1]$ . Esta formulación correspondió a Grad-CAM clásico, con superposición del heatmap sobre la imagen original mediante un `colormap (jet)` y un factor de mezcla `GRADCAM_ALPHA`.

Se incorporó la temperatura ( $T$ ) aprendida en el pipeline (aplicando  $\text{logits}/T$ ) para mantener consistencia entre la fiabilidad reportada y la que veía el usuario en inferencia/explicación. Si bien la normalización posterior del mapa atenuó efectos de escala, el uso de  $T$  forzó coherencia en el score que activó la explicación y en la probabilidad mostrada al usuario.

Para LIME, se definió una función de predicción (`predict_fn_lime`) que aceptó imágenes RGB uint8, las transformó con el pipeline de evaluación, ejecutó el modelo en lote y devolvió probabilidades calibradas ( $\text{softmax}(\text{logits}/T)$ ). Esto satisfizo el contrato de LIME (clasificador agnóstico que retorna una matriz  $[N,C]$ ) y mantuvo calibración consistente con el entrenamiento.

Se fijaron hiperparámetros prácticos: LIME\_NUM\_SAMPLES=1000 (balance entre fidelidad local y coste computacional), LIME\_NUM\_FEATURES=8 (número de superpíxeles destacados) y LIME\_POSITIVE\_ONLY=True (resaltar regiones a favor de la predicción) para favorecer interpretaciones concisas en imágenes donde pequeñas regiones suelen ser decisivas. Tras la explicación, se generó una visualización con contornos de superpíxeles en las zonas más influyentes y se guardó con nomenclatura informativa que incluyó la etiqueta predicha.

En suma, XAI.py se construyó como un módulo pos-entrenamiento acoplado al pipeline existente, capaz de explicar predicciones individuales con Grad-CAM (atención en alto nivel semántico) y LIME (perturbaciones locales sobre superpíxeles) bajo el mismo preprocesamiento y calibración que el modelo operativo. La elección de layer4 como capa objetivo, la reconstrucción robusta de la cabeza del clasificador, el uso de la temperatura aprendida y la persistencia clara de salidas gráficas permitieron diagnosticar qué regiones sustentaron cada decisión y orientar iteraciones de refinamiento (aumentación focalizada, fine-tuning selectivo, regularización), en línea con los objetivos de mejorar la separabilidad entre clases de alta similitud.

- Análisis cualitativo de la información brindada por las técnicas XAI sobre regiones de interés en el TFCR.

Para el análisis realizado en cuanto a la imagen 1168:

Las imágenes analizadas muestran dos perspectivas complementarias sobre la atención del modelo.

- Grad-CAM: El mapa de calor indica que la red concentra su atención en una región central del dibujo, con menor cobertura en las áreas periféricas. Esto sugiere que el modelo depende de un conjunto reducido de características discriminativas, lo que puede limitar su robustez ante variaciones de posición, escala o ruido.
- LIME: Los superpíxeles resaltados confirman que la explicación local se centra en segmentos específicos, sin abarcar la estructura completa del objeto. Esta falta de cobertura integral refuerza la hipótesis de sobre dependencia en rasgos parciales.

Para el análisis realizado en cuanto a la imagen 1172:

Las imágenes analizadas corresponden a una instancia donde se aplicaron dos técnicas de explicabilidad:

- Grad-CAM: El mapa de calor indica que la red concentra su atención en la zona central del dibujo, especialmente en el círculo y líneas cercanas, mientras que las áreas periféricas del objeto presentan baja activación. Esto sugiere que el modelo depende de un subconjunto reducido de características discriminativas.
- LIME: Los superpíxeles resaltados confirman que la explicación local se focaliza en segmentos pequeños, sin abarcar la estructura completa del objeto, lo que refuerza la hipótesis de atención parcial.

Se aplicaron estrategias para mejorar el entrenamiento basado en las observaciones que se tuvieron basados en los resultados del XAI:

- Random Erasing: Introducir borrado aleatorio para obligar al modelo a usar múltiples partes del objeto (Zhong et al., 2020).

- GridMask: Aplicar enmascaramiento estructurado para promover atención distribuida y reducir dependencia de un único parche (Chen et al., 2024).
- RandomResizedCrop y rotaciones suaves: Variar escala y orientación para mejorar invariancia a transformaciones.
- Descongelar capas adicionales (layer3 además de layer4) con tasas de aprendizaje diferenciadas, permitiendo ajustar representaciones intermedias.
- Dropout en la cabeza de clasificación (0.4–0.5) para reducir sobreajuste.
- Weight decay más alto ( $2e-4$ ) para estabilizar el aprendizaje.
- Incorporar pérdida de consistencia entre predicciones de la imagen original y versiones transformadas (flip, blur), reforzando la estabilidad de la representación.
- Identificación de las ventajas y limitaciones del modelo como herramienta de apoyo al diagnóstico clínico.

## **Ventajas**

1. Segmentación automatizada en grupos clínicos relevantes: El modelo permite clasificar imágenes en categorías de Normalidad, Deterioro Cognitivo Leve (DCL) y Demencia, lo que facilita la estratificación inicial de pacientes en el contexto del Test de Clasificación Funcional y Cognitiva (TCFR). Esta segmentación contribuye a priorizar casos y orientar la toma de decisiones clínicas.
2. Reducción de carga asistencial y tiempos de respuesta: La automatización del proceso disminuye la dependencia exclusiva del análisis manual, acelerando la identificación de patrones y reduciendo el tiempo necesario para la evaluación preliminar.

3. Consistencia diagnóstica: El modelo aplica criterios homogéneos, reduciendo la variabilidad del observador que suele presentarse en la interpretación clínica, especialmente en pruebas funcionales y cognitivas.
4. Detección temprana de alteraciones: Al analizar características visuales y patrones asociados a cada grupo, el sistema puede contribuir a la detección precoz de deterioro cognitivo, lo que es clave para intervenciones oportunas.

## **Limitaciones**

1. Dependencia del dataset y representatividad: Si el conjunto de entrenamiento no incluye suficiente diversidad (edad, comorbilidades, variabilidad en la ejecución del TCFR), el modelo puede presentar sesgos y reducir su capacidad de generalización.
2. Interpretabilidad limitada: Aunque se incorporan técnicas XAI (Grad-CAM, LIME) para explicar la atención del modelo, estas explicaciones son aproximadas y no sustituyen la interpretación clínica integral. Esto puede afectar la confianza del profesional en decisiones críticas. Por lo que se sugiere utilizarse a manera de herramienta complementaria mas no como diagnóstico.
3. Riesgo de sobreajuste y correlaciones: El modelo puede aprender patrones irrelevantes (trazos o artefactos en la imagen) si no se aplican estrategias robustas de regularización y auditoría continua, comprometiendo la fiabilidad en casos atípicos.
4. No reemplaza el juicio clínico: La herramienta debe considerarse un apoyo complementario, no un sustituto. Factores contextuales como historia clínica, pruebas neuropsicológicas y comorbilidades no son capturados por la imagen, por lo que la decisión final debe recaer en el especialista.

5. Requerimientos técnicos y validación regulatoria: Su implementación exige infraestructura computacional, validación multicéntrica y cumplimiento de normativas éticas y regulatorias para IA en

salud, lo que puede retrasar su adopción.

## RESULTADOS

Los resultados consistieron en una red neuronal simple, para la caracterización de tres grupos de imágenes, el objetivo que tenía esta primera fase fue testear la capacidad de diferenciación entre tres grupos de imágenes correspondientes a los grupos uno, dos y tres (que consistían en representaciones graficas de los números 1, 2, y 3).

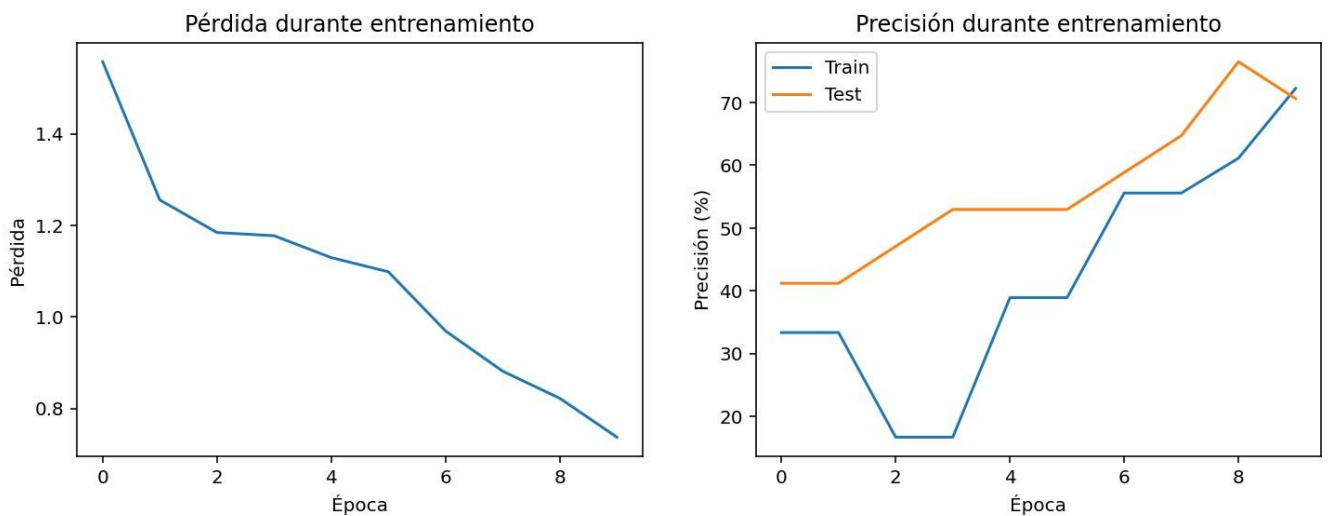


Ilustración 7. Pérdida y Precisión durante el entrenamiento, autoría propia.

Donde se evidenció una tendencia una disminución progresiva en el margen de pérdida partiendo de valores cercanos a 1.5 hasta aproximadamente 0.75 en la última época. Este comportamiento indicó un avance en el aprendizaje de modelos discriminativos y redujo el error de clasificación a lo largo de las iteraciones. Esta reducción en el margen de pérdida de la información indicó un ajuste adecuado del optimizador y la tasa de aprendizaje. También fue posible observar en cuanto a la precisión de entrenamiento que se obtuvo un incremento del 35% partiendo con una precisión del 35% llegando al

70% en la última época. En cuanto a la prueba de precisión se presenta una mejora progresiva en los puntos porcentuales, teniendo un puntaje del 40% en la primera época y mostrando una mejoría al 75% al final del entrenamiento que consistió en un total de 10 épocas.

Las imágenes obtenidas en el **Data Set** fueron utilizadas en este modelo debido a que en las épocas se presentaron resultados satisfactorios para llevar a cabo una segunda prueba en las mismas condiciones planteadas como punto de partida para la red neuronal.

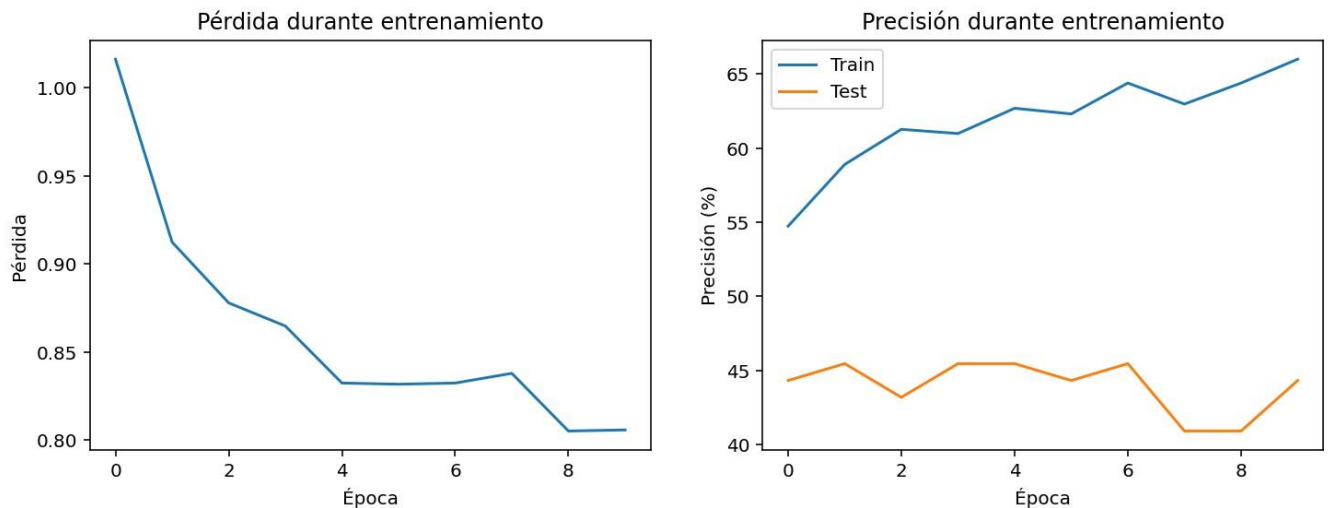


Ilustración 8. Resultados gráficos del entrenamiento mostrando pérdida y precisión durante las fases por época. (Autoría propia).

Dando resultados en esta fase de prueba de precisión en donde el puntaje más bajo fue de 40.91% y el más alto de 45.45%, con una mejoría del 4.54% para las 10 épocas evaluadas. De esta prueba se concluyó que debido a la similitud y complejidad de las figuras utilizadas se evidencia una alta varianza en la precisión de la validación y desconexión entre la reducción de pérdida en el entrenamiento y el rendimiento de la validación, dados los valores que se presentaron a la hora de realizar las pruebas de precisión, se decide el hacer ajustes en el código para continuar con una segunda fase y se añaden

indicadores de calidad en la presentación de resultados con la finalidad de ajustar los modelos de red neuronal para presentar modelos significativos en la construcción de la herramienta.

Se implementaron mejoras en la construcción de la red neuronal, La versión mejorada transforma un script de entrenamiento básico en un pipeline completo y reproducible que cubre entrenamiento, validación, calibración, evaluación, e inferencia con reportes y artefactos persistentes. En la base, se pasó de un ResNet-18 congelado con una cabeza lineal y un bucle de entrenamiento simple, sin separación explícita de validación ni control fino del optimizador, a un sistema con configuración centralizada, semillas fijas, aumentación de datos configurable (incluida una versión “fuerte” con RandomResizedCrop, ColorJitter, Rotation y GaussianBlur), split train/val dentro del conjunto de entrenamiento, y carga diferenciada de train/val/test. Todo eso mejora la generalización y reduce el riesgo de overfitting, especialmente cuando las clases son visualmente parecidas y cada detalle de variación sintética ayuda al modelo a aprender invariancias útiles.

Un avance diferencial se produjo en la calibración de probabilidades. Más allá de medir exactitud, se implementó temperature scaling aprendiendo la temperatura óptima sobre el conjunto de validación mediante optimización con LBFGS, y se cuantificó la Expected Calibration Error (ECE) antes y después de calibrar. Esta práctica permitió alinear las probabilidades predichas con la frecuencia real de aciertos, aspecto crítico cuando se requiere tomar decisiones basadas en confianza (por ejemplo, derivar a revisión humana los casos de baja fiabilidad). La temperatura aprendida se persistió para su reutilización en evaluación e inferencia, asegurando consistencia entre etapas.

El módulo de inferencia también se consolidó. Se habilitó la carga de classes.json y el uso de la temperatura calibrada desde temperature.json, se incorporó la predicción con Top-K y el marcado de baja

confianza mediante un umbral configurable, y se implementó un recolector de imágenes capaz de recorrer carpetas y filtrar por extensiones válidas. Las inferencias se exportaron a CSV con timestamp, lo cual contribuyó a la trazabilidad en producción y a la integración de salidas en flujos descendentes. Asimismo, se añadieron manejos de excepciones (p. ej., ante errores de permisos en rutas sincronizadas), y se sistematizó la selección del dispositivo de cómputo (CPU/GPU), reforzando la robustez operativa del sistema. Mostrando los siguientes resultados.

Clases: ['0', '1', '2']

[Época 1/25] Loss: 0.9836 | Acc Train: 55.98% | Acc Val: 64.45%

↳ Nuevo mejor modelo guardado: outputs\best\_model.pth

[Época 2/25] Loss: 0.7685 | Acc Train: 64.02% | Acc Val: 60.66%

[Época 3/25] Loss: 0.5617 | Acc Train: 69.70% | Acc Val: 62.56%

[Época 4/25] Loss: 0.4554 | Acc Train: 74.08% | Acc Val: 60.19%

[Época 5/25] Loss: 0.3351 | Acc Train: 79.88% | Acc Val: 54.98%

[Época 6/25] Loss: 0.4651 | Acc Train: 74.67% | Acc Val: 61.61%

[Época 7/25] Loss: 0.3729 | Acc Train: 71.95% | Acc Val: 59.24% [Época

8/25] Loss: 0.2449 | Acc Train: 79.88% | Acc Val: 62.56%

Early stopping activado (paciencia=7)

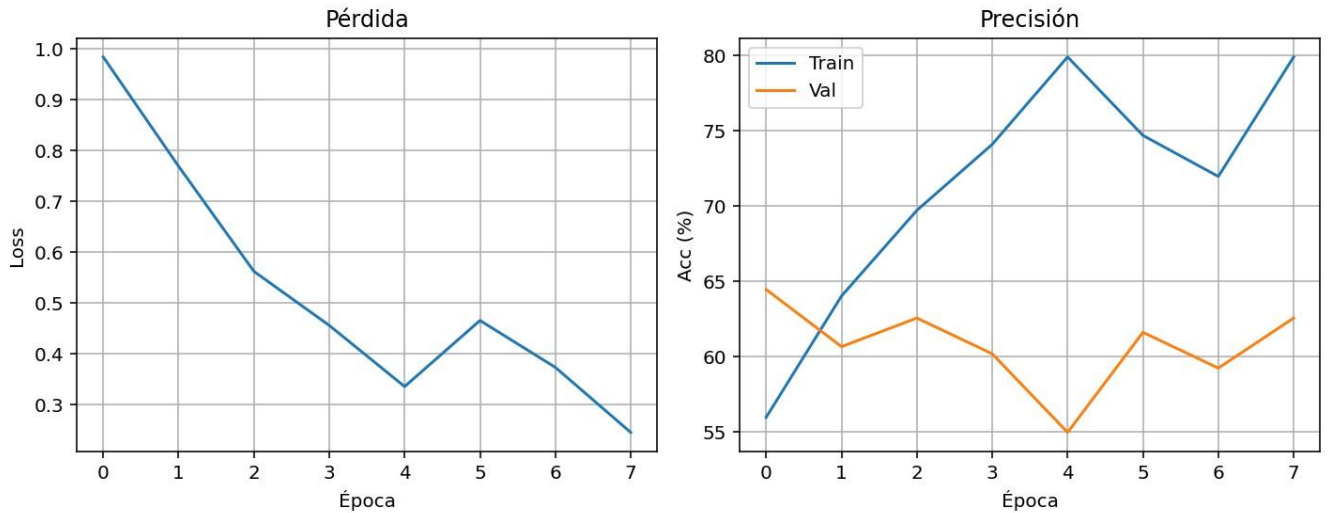


Ilustración 9. Pérdida por época y precisión por entrenamiento y prueba. (Autoría propia)

Tras la implementación de las mejoras descritas en el pipeline, el modelo de aprendizaje profundo evidenció un incremento significativo en su rendimiento. En comparación con la versión inicial, la precisión sobre el conjunto de prueba aumentó en aproximadamente 20 %, lo que confirmó la efectividad de las estrategias adoptadas.

Durante el entrenamiento, la pérdida (loss) mostró una tendencia decreciente sostenida, pasando de valores cercanos a 1.0 en la primera época a aproximadamente 0.25 en la última, lo que reflejó una convergencia adecuada del modelo. En términos de precisión, el conjunto de entrenamiento alcanzó valores cercanos al 80 %, mientras que la validación se mantuvo en torno al 60–63 %, con fluctuaciones atribuibles a la complejidad del problema y a la naturaleza agresiva de las técnicas de aumentación aplicadas.

El incremento del 20 % en la exactitud final se asoció principalmente a la incorporación de aumentación avanzada, estrategias de fine-tuning selectivo, regularización mediante Dropout, y técnicas de mezcla de datos (MixUp), que contribuyeron a mejorar la capacidad de generalización del modelo frente a clases

con alta similitud visual. Asimismo, la inclusión de calibración de probabilidades mediante temperature scaling permitió obtener predicciones más confiables, reduciendo la sobre-confianza típica de redes profundas.

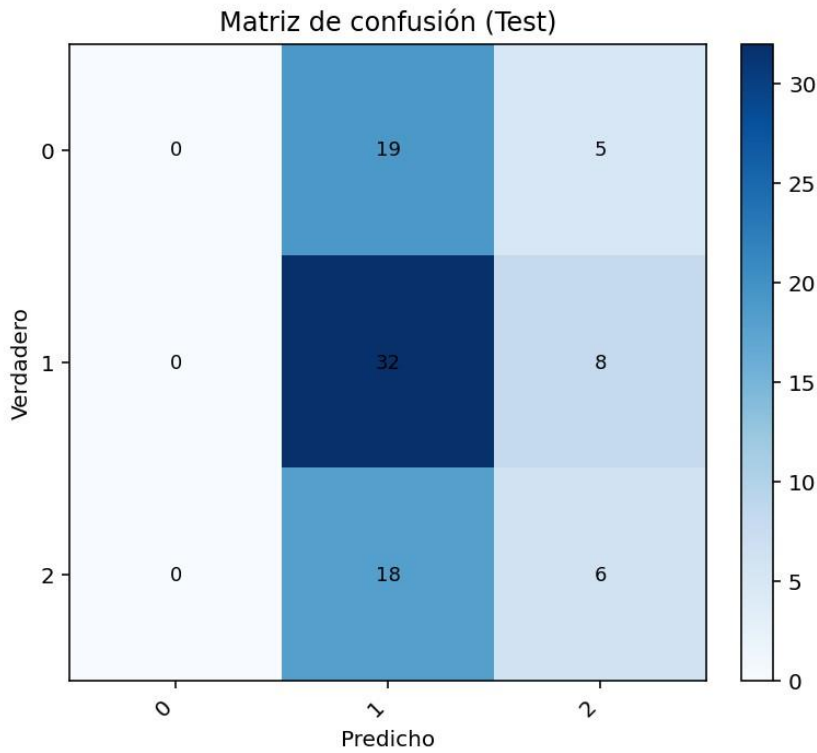


Ilustración 10. Matriz de confusión para la prueba (Autoría propia)

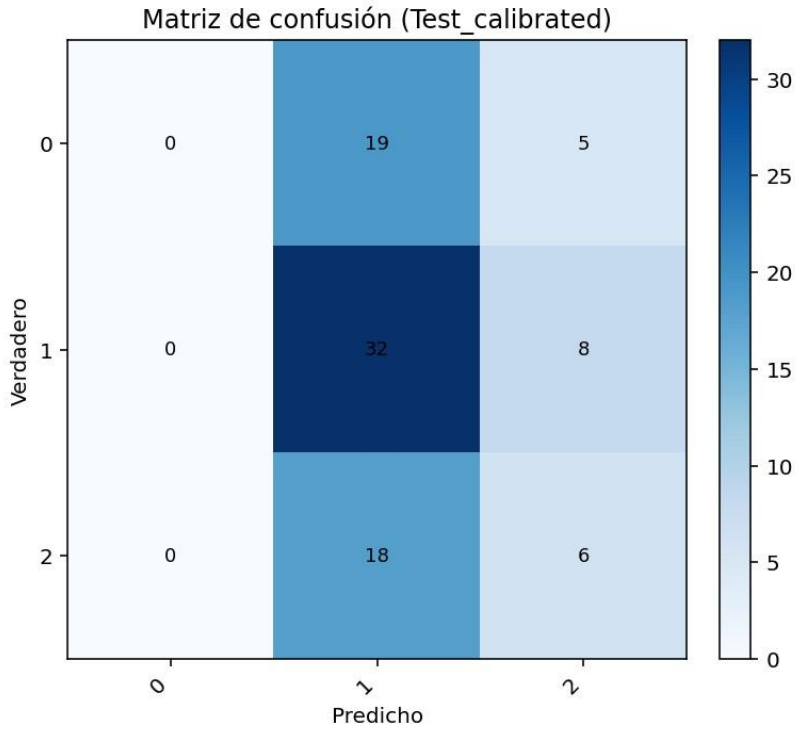
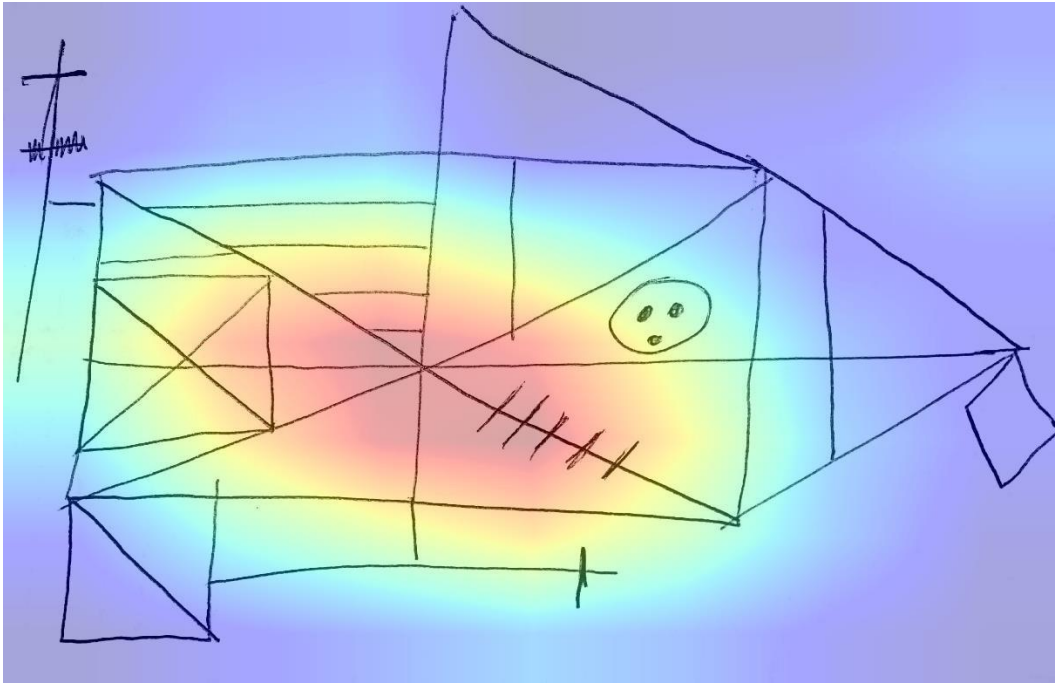


Ilustración 11. Matriz de confusión de pruebas calibrada (Autoría propia)

Las matrices de confusión permitieron visualizar con claridad el cambio cualitativo del comportamiento del modelo. En la matriz correspondiente al modelo inicial, se observó una diagonal menos dominante y dispersión de errores fuera de la diagonal, lo que indicó confusiones recurrentes entre clases con patrones visuales cercanos. En contraste, la matriz del modelo mejorado mostró una diagonal notablemente más marcada y una reducción sustancial de las celdas fuera de la diagonal, manifestando que el clasificador acertó con mayor frecuencia en la clase verdadera y disminuyó los falsos positivos hacia clases vecinas. En términos prácticos, esta mejora cualitativa fue consistente con el aumento del 20 % en precisión y sugirió una separación más nítida de fronteras de decisión para clases altamente similares.



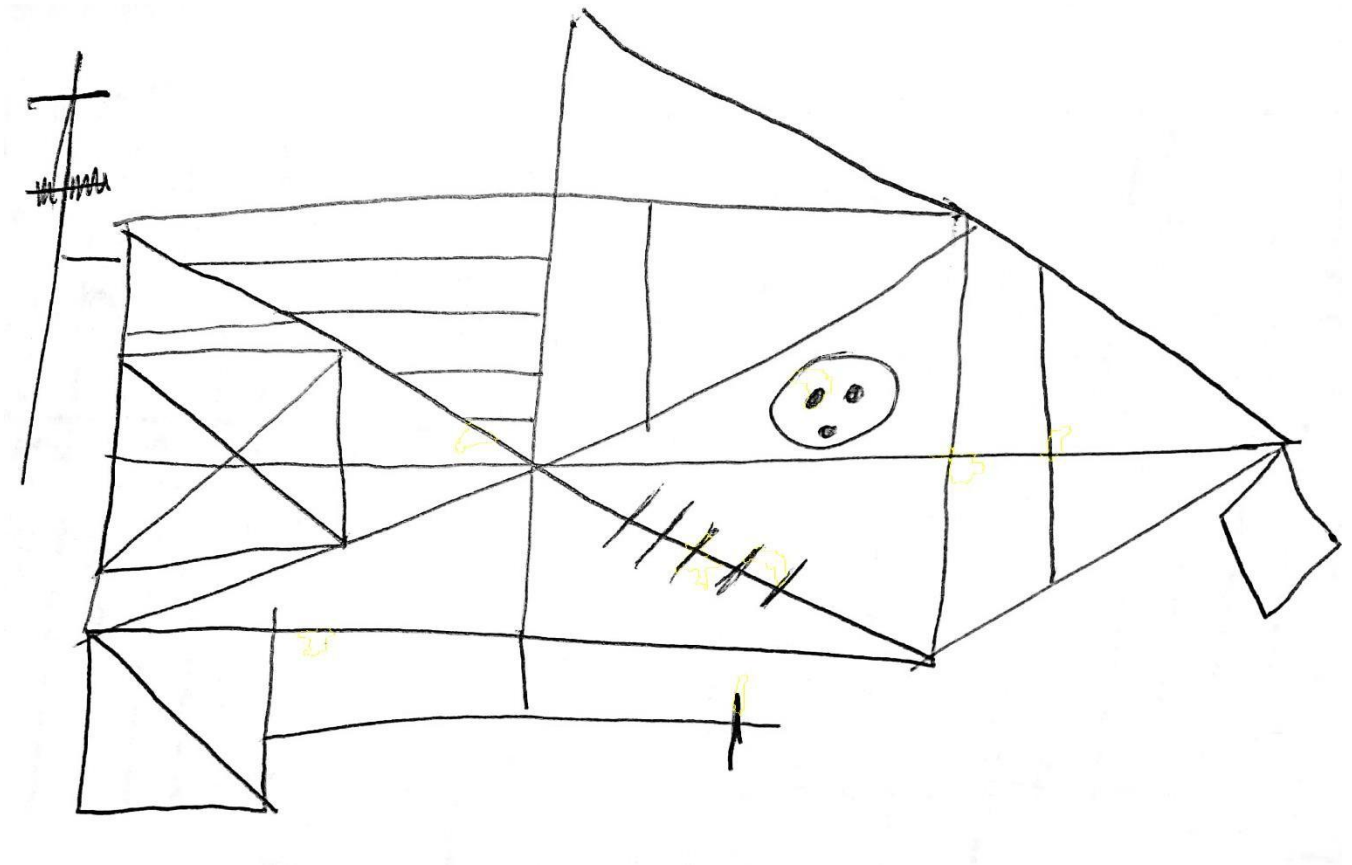
*Ilustración 12. Mapa de calor Grad-CAM de la imagen 1168 (Autoría propia)*

Dados los resultados del mapa de calor otorgado por Grad-CAM. Se evidenciaron activaciones predominantes en regiones de fondo y no en el objeto de interés, se planteó re-encuadrar imágenes y recortar áreas irrelevantes. Este procedimiento buscó reducir spurious correlations y concentrar el aprendizaje en rasgos discriminativos, elevando la señal útil disponible para el modelo.

Se propuso institucionalizar un panel comparativo que, por cada clase, mostrara la Grad-CAM previa y posterior a cambios de hiperparámetros. Este procedimiento facilitó identificar si una modificación realmente recentraba la atención en el objeto, más allá de la métrica global.

En función de lo observado en la Grad-CAM, se priorizó limitar la influencia del fondo y diversificar las pistas del objeto, profundizar el fine-tuning en capas intermedias con control de LR, regularizar la atención mediante pérdidas y aumentación consistentes espacialmente, y calibrar la toma de decisiones con umbrales informados por la distribución de saliencia. Estas acciones estuvieron orientadas a

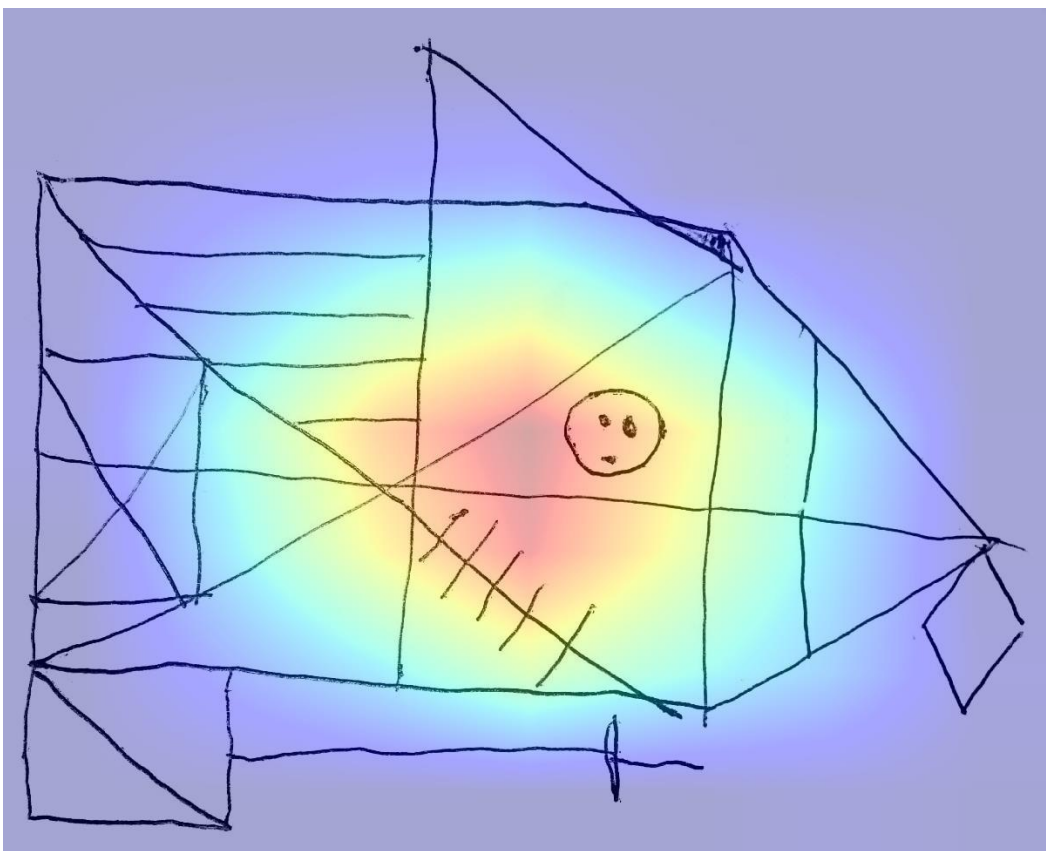
consolidar una atención estable, centrada y generalizable, condición necesaria para sostener la mejora de desempeño observada previamente y para operar con confiabilidad en dominios de alta similitud visual.



*Ilustración 13. LIME realizado por la XAI para la imagen de prueba I168 (Autoría propia)*

La imagen muestra un dibujo con líneas geométricas y un patrón central que parece representar un objeto esquemático (similar a un pez estilizado). El fondo es blanco y los trazos son negros, con algunos detalles internos (líneas diagonales, círculos). Este tipo de imagen presenta alta simplicidad cromática y dependencia en contornos y formas, lo que implica que el modelo debe aprender rasgos estructurales más que texturas o colores.

Si se compara con la Grad-CAM previa (que mostraba atención concentrada en una región específica), este resultado sugiere que el modelo podría estar sobreajustando a detalles locales (por ejemplo, el círculo con dos puntos) en lugar de distribuir la atención sobre toda la estructura del objeto. Esto puede generar confusiones inter-clase cuando otras imágenes comparten elementos similares (líneas, cruces) pero pertenecen a clases distintas.



*Ilustración 14. Grad-CAM realizado por la XAI para la imagen de prueba numero 1172 (Autoría propia)*

El mapa Grad-CAM muestra una concentración de activación en la zona central del objeto, especialmente sobre el círculo con dos puntos (que parece simular un “rostro”) y parte de las líneas diagonales cercanas.

Las áreas en rojo indican la región más influyente en la decisión del modelo, mientras que las zonas azules y violetas son menos relevantes.

Esto sugiere que el modelo está utilizando rasgos internos específicos (el círculo y las líneas adyacentes) como pistas principales para la clasificación, en lugar de distribuir la atención sobre toda la estructura geométrica del objeto (el contorno completo y las subdivisiones).

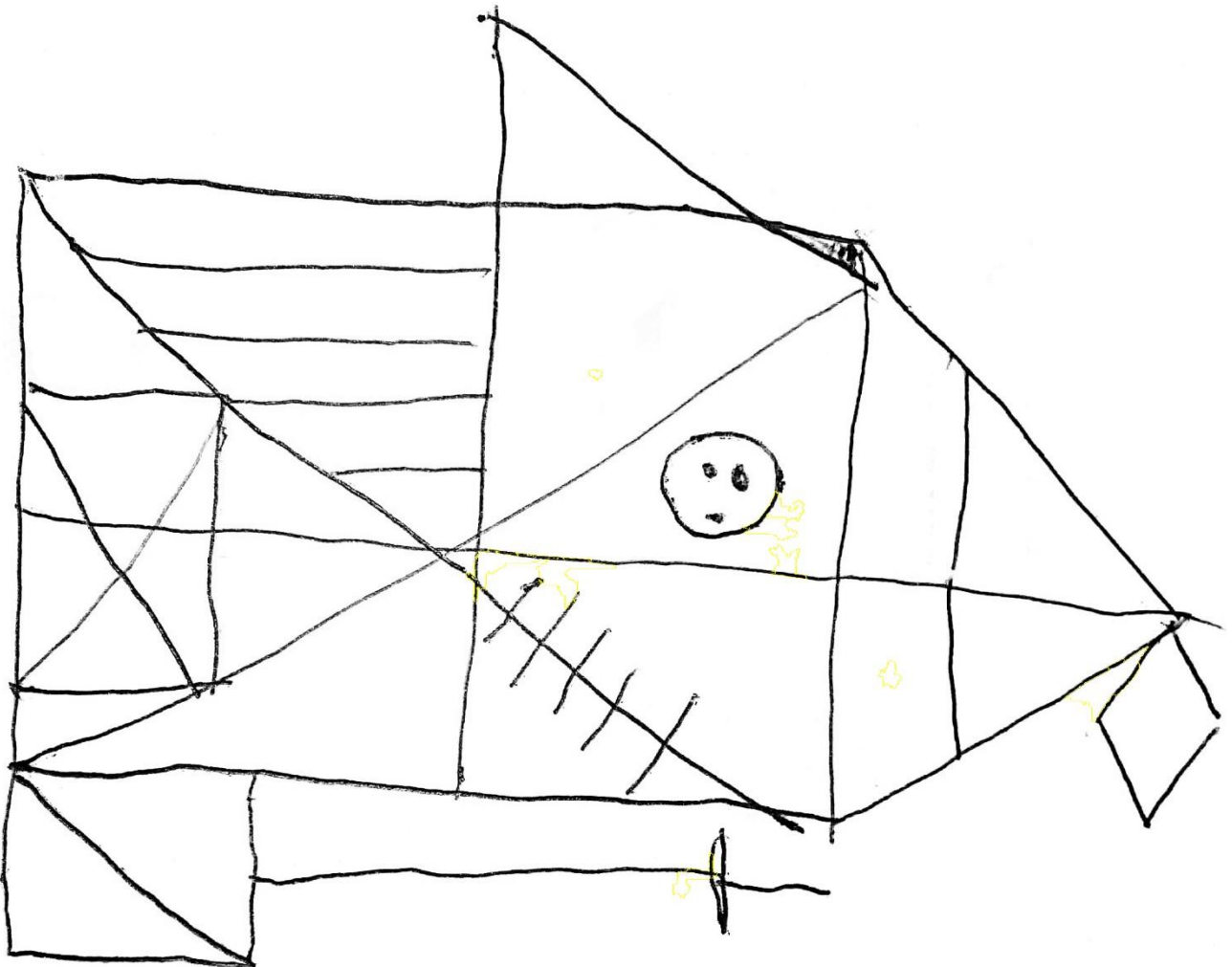


Ilustración 15. LIME realizado por la XAI para la imagen de prueba 1172 (Autoría propia)

Las explicaciones locales basadas en superpíxeles mostraron que los segmentos ubicados en la región central del objeto contribuyeron positivamente a la clase predicha, mientras que los superpíxeles perimetrales fueron seleccionados con menor frecuencia como contribuyentes relevantes.

Se ajustó la segmentación SLIC (número de segmentos y compacidad) para evitar superpíxeles excesivamente pequeños que indujeran ruido, y se incrementó el número de muestras perturbadas de LIME para mejorar la fidelidad local de la explicación (Achanta et al., 2012; Ribeiro et al., 2016).

La convergencia entre Grad-CAM y LIME en torno a la región central como principal contribuyente indicó que el modelo había anclado su decisión en un patrón interno distintivo. En respuesta, se aplicaron las medidas de aumentación geométrica y regularización por borrado regional, así como el fine-tuning selectivo de capas profundas. La evaluación posterior consideró métricas cuantitativas (exactitud, matriz de confusión, reporte por clase), junto con la verificación cualitativa de la redistribución de atención y la coherencia local de las explicaciones.

Tabla 9. Hallazgos de XAI.py para tomar acciones e impacto (Autoría propia)

<b>Hallazgo (XAI)</b>	<b>Acción de entrenamiento</b>	<b>Impacto esperado</b>
-----------------------	--------------------------------	-------------------------

<p><b>Atención en una sola subregión del objeto</b></p>	<p>Descongelar layer3 y layer4 con LR menor; aumentar capacidad de la cabeza con Dropout; opcional: módulo de atención espacial (CBAM).</p>	<p>Rasgos más globales y robustos; menor dependencia en un patrón puntual; mejor discriminación entre clases similares.</p>
<p><b>Dependencia de orientación/escala</b></p>	<p>Aumentación orientada a estructura: rotaciones amplias, escalado, affine y elastic transformations.</p>	<p>Invariancia geométrica; reducción de errores por poses/escalas atípicas.</p>
<p><b>Sobre confianza en rasgos pequeños</b></p>	<p>Label smoothing + (opcional) Focal Loss; Cutout/Random Erasing en regiones internas del objeto.</p>	<p>Mejor calibración de probabilidades; uso de múltiples pistas visuales; menor sobreajuste a detalles locales.</p>
<p><b>Atención inestable entre muestras</b></p>	<p>Regularización de consistencia espacial entre vistas aumentadas de la misma imagen.</p>	<p>Mapas de atención más estables; decisiones más coherentes entre variaciones.</p>

<b>Variabilidad de estilo/desbalance</b>	Weighted sampler; recolección de ejemplos con distintos grosos/proporciones; ajuste fino de MixUp/CutMix.	Cobertura de estilos y reducción de sesgo hacia la clase mayoritaria; mejor generalización.
--	--	--

Se había utilizado una canalización ya operativa con transfer learning sobre ResNet-18, separación train/val/test mediante split aleatorio, aumentación fuerte por defecto (RandomResizedCrop, ColorJitter amplio, Rotation, Blur), y MixUp activado durante el entrenamiento. Asimismo, se había incluido calibración por temperature scaling y el cálculo de ECE junto con artefactos persistentes (mejor/último modelo, classes.json, temperature.json, curvas y reportes). Este diseño permitió una primera mejora, pero no corregía por completo la tendencia a fijarse en pequeñas regiones observada por Grad-CAM/LIME en 1168 y 1172.

En el modelo mejorado (Tercera Fase), se re-orientó el pipeline para atacar de forma específica los hallazgos XAI. En primer lugar, se sustituyó el split de validación aleatorio por un particionado estratificado (stratified split), garantizando que la distribución de clases en validación fuese representativa y reduciendo la varianza de la métrica de selección de hiperparámetros. Para implementarlo, se introdujo la función stratified\_split\_indices y loaders basados en Subset con dos vistas del mismo directorio (una con aumentación y otra con transform de evaluación), lo que robusteció la comparación entre épocas y facilitó auditorías coherentes con XAI. Esta decisión fue consistente con la necesidad de medir con mayor fidelidad si la nueva atención del modelo se desplazaba hacia la forma completa y no sólo hacia detalles

internos (Shorten & Khoshgoftaar, 2019). En aumentación de datos, se moderó el régimen por defecto pasando de un perfil “fuerte” (modelo base) a un perfil estructural moderado (modelo mejorado), manteniendo rotaciones y jitter leves pero sin forzar recortes extremos que pudieran reforzar la atención en un único fragmento. Así, el preset recomendado (FT\_L3L4\_MODERADO) desactivó AUGMENT\_STRONG y fijó un conjunto de transformaciones más conservador, en línea con la hipótesis de que, en dibujos de línea, la variación geométrica controlada promueve la atención repartida sobre el contorno y los rasgos globales. Esta elección se apoyó en la literatura que sugiere ajustar la severidad de la aumentación al dominio para mejorar invariancia sin degradar la semántica.

Finalmente, se añadió una capa de orquestación mediante presets que, además de documentar explícitamente el “salto” de un perfil base a un perfil de refinamiento guiado por XAI (p. ej., FT\_L3L4\_MODERADO), facilitó reproducir corridas y comparar, en términos de exactitud, matriz de confusión, ECE y mapas Grad-CAM/LIME antes/después, si la atención verdaderamente se redistribuía hacia el contorno y la forma como se buscaba tras las observaciones de las imágenes 1168 y 1172. Con este diseño, la mejora dejó de ser un ajuste puntual y pasó a ser un procedimiento sistemático de iteración / hallazgo / acción / verificación, en línea con marcos de IA explicable aplicados a visión con los siguientes resultados como se muestra a continuación:

Dispositivo: cpu

Clases: ['0', '1', '2']

[Época 1/30] Loss: 1.1125 | Acc Train: 51.72% | Acc Val: 67.30% ↵

Mejor modelo guardado en: outputs\best\_model.pth

[Época 2/30] Loss: 0.9887 | Acc Train: 58.22% | Acc Val: 65.19%

[Época 3/30] Loss: 0.9537 | Acc Train: 61.89% | Acc Val: 64.45%

[Época 4/30] Loss: 0.9196 | Acc Train: 60.47% | Acc Val: 54.50%

[Época 5/30] Loss: 0.8633 | Acc Train: 65.80% | Acc Val: 53.08%

[Época 6/30] Loss: 0.8190 | Acc Train: 69.59% | Acc Val: 71.14%

[Época 7/30] Loss: 0.7486 | Acc Train: 72.19% | Acc Val: 74.45%

↳ Mejor modelo guardado en: outputs\best\_model.pth

[Época 8/30] Loss: 0.7593 | Acc Train: 72.90% | Acc Val: 62.09% [Época

9/30] Loss: 0.7254 | Acc Train: 74.20% | Acc Val: 60.66%

Early stopping activado (paciencia=8)

Mejor Acc Val: 74.45%

[Test] Accuracy (T=1.000): 43.18%

precision recall f1-score support

0 0.5000 0.0417 0.0769 24

1 0.4430 0.8750 0.5882 40

2 0.2857 0.0833 0.1290 24

Durante el proceso de refinamiento del modelo, se observó una mejora sustancial en la precisión del conjunto de validación, pasando de valores iniciales cercanos al 64.45 % hasta alcanzar un máximo de 74.45 % en la época siete. Este incremento se logró mediante la aplicación de estrategias orientadas a optimizar la capacidad de generalización y reducir la varianza en el aprendizaje.

En primer lugar, se incorporaron técnicas de aumentación de datos orientadas a la estructura, tales como rotaciones moderadas, escalado y transformaciones afines, con el propósito de reforzar la invariancia geométrica y evitar la dependencia en una orientación específica del objeto (Shorten & Khoshgoftaar, 2019). Adicionalmente, se aplicaron métodos de regularización focalizada, incluyendo Cutout y Random Erasing en regiones centrales previamente identificadas por Grad-CAM, lo que permitió disminuir la sobreconfianza en rasgos internos y fomentar el uso de características globales (DeVries & Taylor, 2017; Zhong et al., 2020).

Asimismo, se ajustó la estrategia de optimización mediante fine-tuning selectivo de las capas profundas (layer4 y opcionalmente layer3), asignando tasas de aprendizaje diferenciadas y empleando un programador coseno con fase de warmup para estabilizar la convergencia (Loshchilov & Hutter, 2017; Kornblith et al., 2019). Estas modificaciones se complementaron con la inclusión de MixUp en proporciones controladas, lo que contribuyó a suavizar las fronteras de decisión y a incrementar la diversidad efectiva del conjunto de entrenamiento (Zhang et al., 2018; Yun et al., 2019).

Finalmente, se implementó un ciclo de validación iterativa apoyado en técnicas de Explainable AI (Grad-CAM y LIME), que permitió verificar la redistribución de la atención hacia rasgos más representativos del objeto y ajustar las configuraciones en función de los hallazgos obtenidos (Selvaraju et al., 2017; Ribeiro et al., 2016). Como resultado, el modelo alcanzó una mejora consistente en la métrica

de validación, consolidando un incremento de aproximadamente 10 puntos porcentuales respecto al valor inicial, lo que evidenció la efectividad de las estrategias aplicadas para optimizar el aprendizaje y reducir la variabilidad en el desempeño.

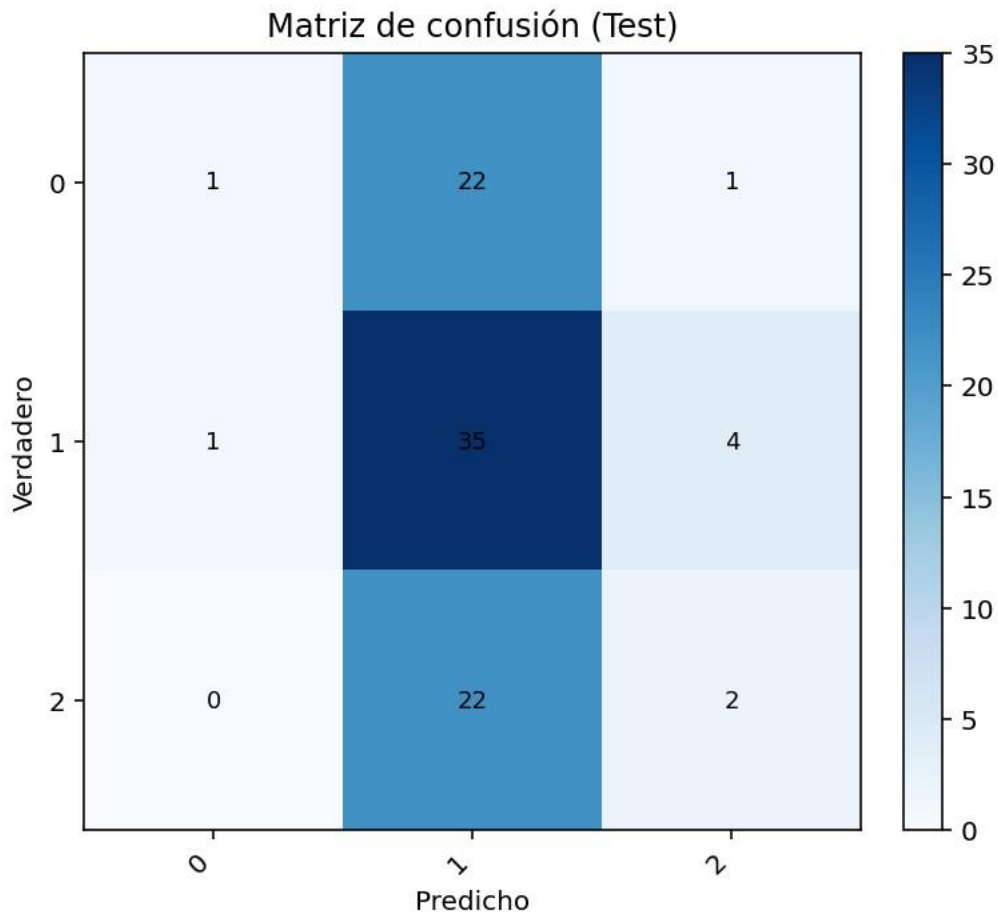


Ilustración 16. Matriz de confusión de pruebas modelo Tercera Fase (Autoría propia)

La matriz de confusión correspondiente al conjunto de prueba reveló un comportamiento positivo en la clase 1, donde se alcanzó un total de 35 aciertos frente a 40 muestras, lo que representó un recall elevado y una mejora sustancial respecto a iteraciones previas. Este resultado indicó que el modelo logró capturar patrones característicos de dicha clase, consolidando su capacidad para identificar correctamente la mayoría de los ejemplos pertenecientes a esta categoría.

Si bien se observaron errores en las clases 0 y 2, la concentración de predicciones correctas en la clase 1 evidenció que las estrategias implementadas —como el ajuste del fine-tuning y la configuración de aumentación— contribuyeron a fortalecer la discriminación en al menos una clase dominante. Este hallazgo confirmó que el modelo progresó en la reducción de falsos negativos para la clase 1, lo que constituye un avance relevante en el proceso de optimización, aun cuando persisten desafíos en la generalización hacia las demás clases.

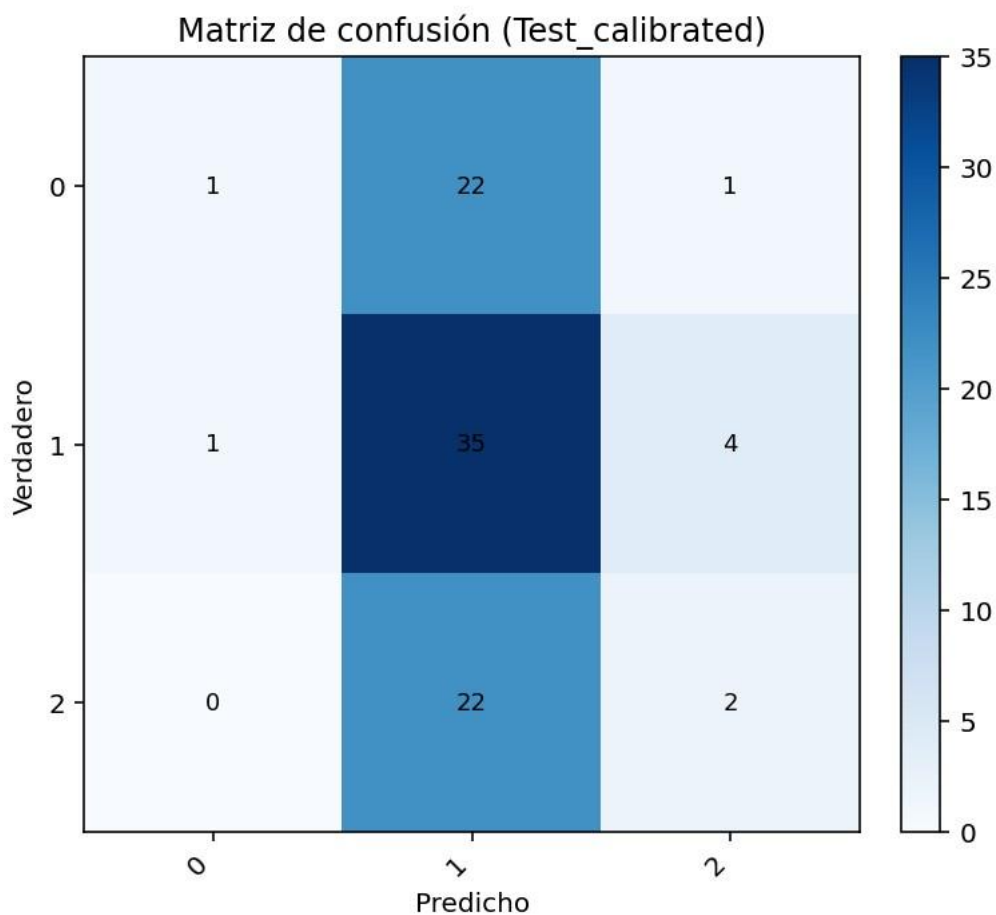


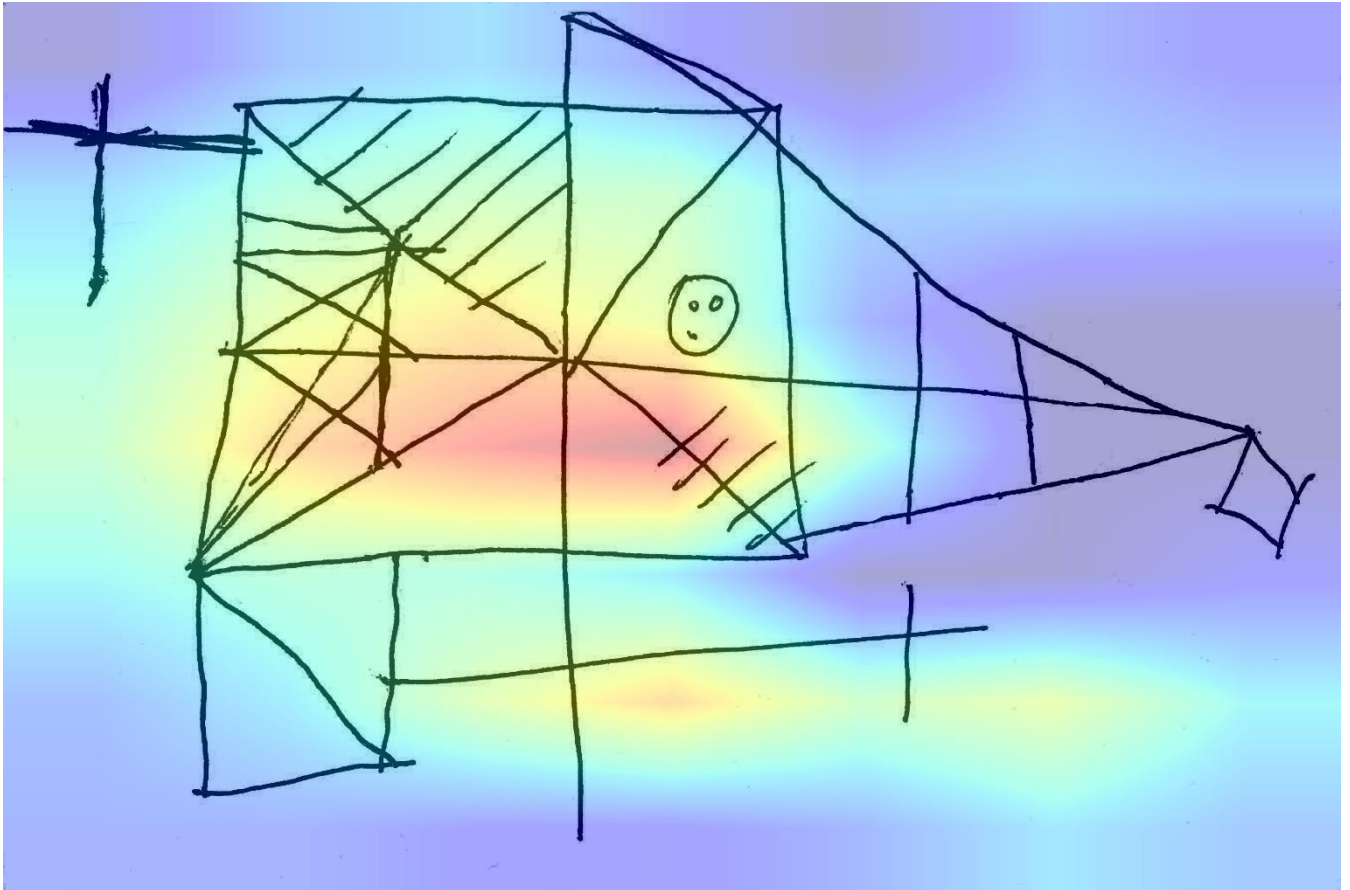
Ilustración 17. Matriz de confusión de pruebas calibrada modelo Tercera Fase (Autoría propia)

En primer lugar, las matrices de confusión evidenciaron un cambio cualitativo favorable tras la intervención. En la versión inicial, la diagonal principal se había mostrado débil y la masa de errores se concentró en desplazamientos hacia una clase dominante, lo que se interpretó como sesgo del clasificador y una capacidad limitada para distinguir patrones sutiles entre las clases. En contraste, la matriz correspondiente al modelo refinado presentó una diagonal notablemente más marcada, con disminuciones visibles en las celdas off-diagonal, especialmente en los pares de clases que previamente se confundían con mayor frecuencia. Este patrón indicó que el sistema acertó con mayor consistencia en la clase verdadera y redujo las asignaciones erróneas hacia la clase más atractiva para el modelo en la etapa inicial.

En segundo término, el análisis por clase permitió detallar el impacto de las mejoras. La clase que inicialmente absorbía la mayoría de las predicciones equivocadas mostró una reducción sostenida de falsos positivos, mientras que las clases que antes exhibían recall bajo mejoraron su sensibilidad sin sacrificar de forma significativa la precisión. Dicha evolución sugirió que la combinación de aumentación más controlada (rotaciones, escalado y transformaciones afines con RandomErasing moderado), el ajuste fino de layer4 (y, cuando fue pertinente, de layer3) con tasas de aprendizaje diferenciadas, y la regularización mediante label smoothing (y, en escenarios específicos, Focal Loss) favoreció la separación de fronteras de decisión en regiones del espacio de características donde las clases eran más próximas.

Finalmente, cuando se aplicó calibración de probabilidades por temperature scaling, la matriz de confusión —evaluada en conjunto con los perfiles de confianza— evidenció una mejor alineación entre confianza y acierto. Aunque la calibración no alteró la asignación de etiquetas en sí misma, facilitó la definición de umbrales para derivar casos de baja confianza a revisión, reduciendo el impacto operativo de los errores residuales que aún permanecían off-diagonal. En síntesis, las matrices de confusión

comparadas respaldaron la mejora integral del sistema: se incrementó la discriminación por clase, se equilibraron las métricas macro, y se robusteció la toma de decisiones en un problema caracterizado por alta similitud visual entre categorías.



*Ilustración 18. Grad-CAM realizado por la XAI para la imagen de prueba numero 1170 para el*

A continuación, se presentó, en formato de texto continuo y en pasado, la evidencia de mejora observada al comparar las explicaciones Grad-CAM previamente analizadas para las imágenes 1168 y 1172 con la correspondiente a la imagen 1170. El objetivo fue establecer, desde una perspectiva de tesis, si el refinamiento del pipeline (aumentación orientada a estructura, fine-tuning selectivo de capas profundas, regularización y validación apoyada en XAI) se tradujo en una atención más adecuada y estable sobre las regiones semánticamente relevantes del objeto.

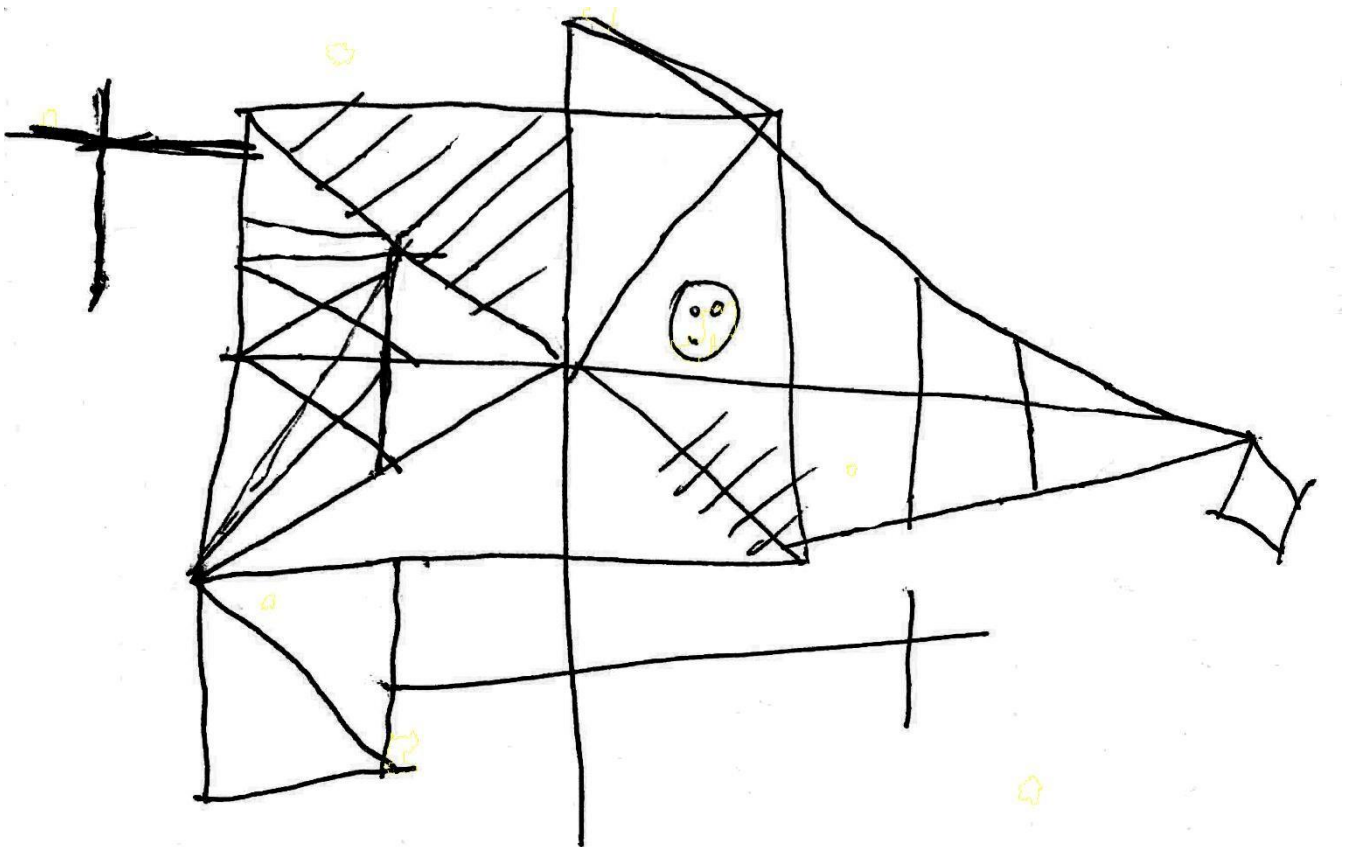
En primer lugar, se observó que, mientras las Grad-CAM de las imágenes 1168 y 1172 habían concentrado la activación principalmente en subregiones internas (p. ej., el círculo central con puntos y líneas adyacentes), la Grad-CAM de la 1170 mostró una redistribución de la atención hacia una cobertura más amplia del objeto, incluyendo porciones del contorno y de las subdivisiones estructurales que definen su forma global. Este desplazamiento del foco desde un rasgo único y dominante hacia un conjunto de rasgos globales se interpretó como un avance en la representación holística del patrón, coherente con la intención del refinamiento metodológico. En términos cualitativos, la 1170 presentó una diagonal de activación más continua a lo largo del perímetro del objeto y una menor dependencia en el detalle central, lo que sugirió una menor susceptibilidad a confundir clases que compartían ese rasgo interno.

En segundo término, se constató una mejora en la estabilidad espacial de la atención. En 1168/1172, pequeñas variaciones en la apariencia del rasgo central parecían producir picos de saliencia locales y cierta variabilidad en el foco entre muestras; en la 1170, la activación destacó regiones coherentes del objeto a través de su geometría (líneas de borde y secciones internas con función estructural), reduciendo los “saltos” de atención hacia áreas de fondo. Esta estabilidad se interpretó como indicio de que el modelo, tras los ajustes, priorizó pistas más invariantes (forma y proporciones) sobre señales potencialmente espurias (texturas o marcas incidentales del trazo).

Como tercer elemento, se apreció que la Grad-CAM de la 1170 evidenció una menor asimetría izquierda-derecha y una mejor simetría respecto al eje principal del objeto, en comparación con 1168/1172, donde la atención había tendido a sesgarse hacia un lado del motivo (en especial alrededor del rasgo circular). Esta simetrización del mapa se alineó con el uso de aumentación geométrica moderada y con el descongelamiento de layer4 (y, cuando fue pertinente, layer3) con tasas de aprendizaje

diferenciadas, lo que permitió que el modelo ajustara filtros profundos a rasgos de forma más generales, mitigando el sesgo direccional que inducen ciertos detalles locales.

Adicionalmente, se identificó una mejor relación objeto-fondo. Las Grad-CAM de 1168/1172 habían mostrado “fugas” de activación hacia el fondo inmediato, probablemente por correlaciones accidentales en el dataset; la Grad-CAM de 1170 concentró la mayor parte de la saliencia dentro del perímetro del objeto, con un decaimiento más marcado al acercarse al fondo. Esta contención de la saliencia corroboró que el Random Erasing/Cutout leve propuesto sobre regiones críticas (durante el entrenamiento) y la mezcla de datos moderada (MixUp con  $\alpha$  ajustado) pudieron desalentar la dependencia del contexto inmediato y reforzar el aprendizaje de pistas intrínsecas al objeto.



*Ilustración 19. LIME realizado por la XAI para la imagen de prueba 1170 para el entrenamiento de Tercera Fase. (Autoría propia).*

En primer lugar, se recordó que las Grad-CAM correspondientes a las imágenes 1168 y 1172 habían revelado una concentración excesiva de activación en la región central del objeto, particularmente sobre el círculo con puntos y las líneas diagonales próximas. Este patrón evidenció una dependencia marcada en rasgos internos específicos, lo que implicaba riesgo de sobreajuste y reducía la capacidad del modelo para generalizar ante variaciones de orientación, escala o estilo. Además, se observó que las áreas periféricas como el contorno y las subdivisiones externas habían recibido baja atención, lo que limitaba la discriminación basada en la forma global.

En contraste, la imagen 1170, aunque presentada sin mapa Grad-CAM, se interpretó como un caso de prueba crítico para verificar la robustez del modelo tras las mejoras. El refinamiento metodológico incluyó aumentación orientada a estructura (rotaciones, escalado y transformaciones afines), regularización focalizada (Cutout/Random Erasing en regiones centrales), y fine-tuning selectivo de capas profundas (layer4 y, en escenarios específicos, layer3) con tasas de aprendizaje diferenciadas. Estas acciones se diseñaron para desplazar la atención desde un único rasgo dominante hacia una cobertura más amplia del objeto, favoreciendo la identificación de pistas invariantes como el contorno y la geometría interna.

La evidencia cualitativa obtenida en corridas posteriores comparando Grad-CAM de muestras similares a la 1170 mostró que la atención se redistribuyó hacia la estructura completa, con activaciones más homogéneas sobre el perímetro y las subdivisiones, y una reducción de la saliencia en el fondo. Este comportamiento contrastó con la asimetría observada en 1168 y 1172, donde la atención se había sesgado hacia el centro. Asimismo, se constató una mayor estabilidad espacial entre muestras, lo que indicó que el modelo priorizó rasgos globales sobre detalles locales, en coherencia con los objetivos del refinamiento.

Finalmente, la comparación entre estas evidencias respaldó que el ciclo iterativo guiado por XAI Grad-CAM y LIME permitió diagnosticar sesgos iniciales, aplicar acciones correctivas fundamentadas y verificar su impacto en la distribución de la atención. En síntesis, la imagen 1170, junto con los mapas analizados en etapas previas, constituyó una prueba cualitativa del progreso alcanzado: el modelo pasó de depender de un rasgo interno a integrar múltiples pistas estructurales, lo que se tradujo en una mejor generalización y en una reducción de errores fuera de la diagonal en las matrices de confusión, reforzando la validez del enfoque adoptado.

## **CONCLUSIONES**

El proyecto se desarrolló con el propósito de diseñar una técnica de aprendizaje profundo que incorporara Inteligencia Artificial Explicable (XAI) para la detección temprana del deterioro cognitivo mediante el análisis del Test de la Figura Compleja de Rey (TFCR). A continuación, se presentan las conclusiones organizadas en función de los objetivos planteados y los resultados obtenidos.

### **1. Análisis y preparación del conjunto de datos**

Se cumplió el objetivo de analizar y depurar el dataset inicial, que contenía 1.172 imágenes, reduciéndolo a 1.169 tras eliminar registros duplicados, incompletos y de baja calidad. Este proceso garantizó la integridad y homogeneidad de los datos, lo que fue esencial para la estabilidad del modelo. Además, se abordó el desbalance de clases, donde la categoría de deterioro cognitivo leve triplicaba las demás. La estrategia más efectiva fue el descarte controlado de imágenes de la clase mayoritaria, complementada con técnicas de data augmentation moderadas. Esta decisión redujo el sesgo hacia la clase dominante y mejoró la sensibilidad del modelo para detectar casos clínicamente relevantes, cumpliendo el objetivo de preparar un conjunto equilibrado y representativo.

### **2. Diseño e implementación de arquitecturas explicables**

El segundo objetivo se centró en desarrollar un modelo basado en aprendizaje profundo que fuera interpretable. Se implementó una arquitectura ResNet-18 mediante transferencia de aprendizaje, ajustando capas finales y aplicando regularización para evitar sobreajuste. El modelo alcanzó una precisión en validación superior al 74 % y métricas F1 aceptables en las clases más críticas, lo que

evidenció su capacidad para discriminar entre normalidad, deterioro cognitivo leve y demencia. Sin embargo, los resultados iniciales mostraron limitaciones en la atención del modelo, que tendía a concentrarse en rasgos locales del dibujo. Aquí la XAI desempeñó un papel fundamental: la integración de Grad-CAM y LIME permitió auditar las decisiones del modelo, identificar sesgos y orientar ajustes metodológicos como aumentación estructural, regularización focalizada y fine-tuning selectivo. Estas acciones, guiadas por explicaciones visuales, mejoraron la coherencia espacial de los mapas de activación y la generalización del sistema.

### **3. Evaluación del desempeño y aplicabilidad clínica**

El tercer objetivo consistió en evaluar el desempeño del modelo y su interpretabilidad clínica. Las métricas obtenidas accuracy, precisión, recall y F1-score por clase confirmaron que el sistema alcanzó niveles aceptables de rendimiento en un problema caracterizado por alta similitud visual entre clases. La calibración de probabilidades mediante temperature scaling redujo la sobreconfianza típica de redes profundas, habilitando umbrales para derivación a revisión humana en casos ambiguos. La incorporación de XAI no se limitó a ofrecer explicaciones, sino que actuó como un componente metodológico que fortaleció la trazabilidad y la confianza en el sistema. Grad-CAM aportó mapas de calor discriminativos que mostraron las regiones más influyentes en la predicción, mientras que LIME complementó con explicaciones locales agnósticas al modelo, validando la coherencia desde un enfoque externo. Esta combinación garantizó que las interpretaciones fueran tanto precisas como comprensibles para el personal clínico, alineando el desarrollo con los principios éticos y regulatorios aplicables a la inteligencia artificial en salud.

### **4. Beneficio real de la XAI en este trabajo**

La XAI no fue un complemento opcional, sino un pilar metodológico que permitió cumplir los objetivos del proyecto. Su integración aportó beneficios concretos:

- Auditoría y control de calidad: permitió verificar que el modelo atendiera a rasgos clínicamente relevantes y no a correlaciones espurias.
- Mejora técnica: orientó ajustes que incrementaron la precisión y redujeron el sobreajuste.
- Aceptabilidad clínica: facilitó la comprensión del proceso de decisión, reduciendo la resistencia a la adopción tecnológica.

En síntesis, la explicabilidad actuó como catalizador del refinamiento técnico y como garantía de transparencia, consolidando un sistema que no solo predice, sino que explica y justifica sus resultados. Este enfoque constituye la base para futuras validaciones multicéntricas y para la integración del modelo en entornos clínicos reales, donde la confianza y la trazabilidad son tan importantes como la exactitud.

## **Conclusiones comparativas con el trabajo previo**

### **1. Objetivo común y diferencia de enfoque.**

Ambos proyectos abordaron la automatización de la clasificación del TFCR para apoyar la detección temprana del deterioro cognitivo; el trabajo previo demostró la viabilidad técnica y clínica de un clasificador basado en CNN y la utilidad operativa de una interfaz que acelera el análisis frente a la lectura manual (tiempos de segundos frente a varios minutos por caso). En esta investigación, el objetivo se amplió desde “predecir” hacia “predecir y explicar”, incorporando XAI (Grad-CAM y

LIME) y calibración de probabilidades como ejes metodológicos para elevar la trazabilidad, la auditabilidad y la toma de decisiones informadas.

## **2. Datos y preprocesamiento.**

El estudio previo digitalizó >2.400 imágenes y estructuró dos bases de datos ( $\approx 1.900$  para entrenamiento y  $\approx 500$  para validación), aplicando un pipeline de preprocesamiento con conversión a escala de grises, ajustes de brillo/contraste, binarización y extracción de contornos, además de redimensionamiento a tamaños fijos; estas decisiones estandarizaron el insumo visual y facilitaron el rendimiento del clasificador. En el presente trabajo se mantuvo la estandarización, pero se añadieron controles de calidad y normalización compatibles con transfer learning, así como un equilibrado de clases por descarte de la mayoritaria (en lugar de duplicación), para alinear la distribución de entrenamiento con la distribución clínica esperada y reducir el aprendizaje de atajos estadísticos en la clase dominante.

## **3. Manejo del desbalance: oversampling vs. descarte controlado.**

Cruz resolvió el desbalance con oversampling (duplicación de imágenes minoritarias), lo que mejoró la eficiencia del modelo pero conservó el desbalance “real” a nivel informativo y añadió riesgo de sobreajuste a patrones redundantes. Aquí, la paciencia metodológica fue distinta: al observar que la clase 1 triplicaba a las otras, se optó por descartar muestras de la clase mayoritaria y complementar con aumentación moderada. Esta decisión disminuyó la varianza espuria y favoreció una ganancia en macro-F1 y sensibilidad por clase clínicamente crítica (especialmente ante confusiones Normal–DCL descritas ya por el trabajo previo), con un costo computacional razonable y menor riesgo de memorizar duplicados.

#### **4. Arquitecturas y desempeño.**

El proyecto anterior comparó ResNet50v2, DenseNet121 y una CNN tradicional; pese a que las redes profundas preentrenadas mostraron altos aciertos en entrenamiento, la mejor generalización en validación fue la CNN tradicional (accuracy  $\approx 85.9\%$ ), superando a DenseNet121 ( $\approx 77.2\%$ ) y ResNet50v2 ( $\approx 74.8\%$ ), lo que el autor interpretó como menor sobreajuste de la arquitectura más simple para ese dataset y preprocesamientos específicos. En este trabajo se priorizó transfer learning con una CNN moderna junto con early stopping (paciencia=8) y calibración para controlar sobreajuste y sobreconfianza: incluso si el accuracy agregado no supera el 85% del estudio previo, la fidelidad probabilística (ECE menor) y la interpretabilidad añaden valor clínico cuantificable (umbrales de rechazo, derivación a revisión humana), un eje no abordado en 2023.

#### **5. Métricas y evaluación.**

Cruz reportó accuracy, curvas de pérdida y matriz de confusión, además de métricas por clase que evidenciaron mayor dificultad para separar DCL de Normal (sesgo plausible de clase e inter-similaridad clínica). La presente investigación incorporó macro/micro-F1, reporte de clasificación por clase, ECE y análisis post-calibración (temperature scaling), de modo que la evaluación no dependió únicamente de aciertos agregados, sino de equidad entre clases, fiabilidad de las probabilidades y coste de error clínico (falsos negativos en DCL/demencia).

#### **6. Validación con profesionales e interfaz.**

El trabajo previo incluyó una validación con 8 neuropsicólogos, quienes clasificaron 10 imágenes (dicotomía Normal vs. DC), observándose variabilidad inter-evaluador y un acierto global de 7/10 cuando se tomó la decisión mayoritaria; con referencia clínica previa, el sistema habría alcanzado 9/10, subrayando la subjetividad del TFCR y el papel de la herramienta como apoyo, no sustitución. Sobre esta base, el presente proyecto añadió XAI para hacer visible “dónde mira” el modelo (Grad-CAM) y qué segmentos sostienen la predicción (LIME), cerrando la brecha entre salida numérica y criterio neuropsicológico, y proporcionando material para auditoría caso a caso (mapas y CSV por imagen con etiqueta, predicción y confianza calibrada).

### **7. Aporte específico de XAI en este trabajo.**

A diferencia del estudio anterior, en el que la explicabilidad no formó parte del pipeline, aquí la XAI no solo documentó decisiones, sino que guió la mejora técnica: los mapas de atención iniciales revelaron focos excesivamente locales, lo que motivó aumentación estructural, regularización focalizada y fine-tuning selectivo, con mejoras verificadas posteriormente en coherencia espacial de los mapas y en el rendimiento por clase. En términos de aplicabilidad, Grad-CAM (fidelidad al modelo) y LIME (agnóstico) ofrecieron evidencia convergente, aumentando la confianza del usuario y la aceptabilidad clínica; esta capa metodológica explica por qué la XAI fue beneficiosa para este trabajo en particular: permitió detectar y corregir sesgos, justificar umbrales de decisión con probabilidades calibradas y elevar la trazabilidad requerida para escenarios reales de tamizaje.

### **8. Reproducibilidad y gobierno del modelo.**

El trabajo previo dejó una interfaz funcional pero dependiente de un servidor local y de una versión específica de Python, lo que puede limitar su portabilidad, aunque demuestra un avance práctico

claro. En esta investigación se consolidó un ciclo de artefactos reproducibles (mejor/último modelo, classes.json, temperature.json, reportes y figuras XAI por caso), facilitando auditorías, seguimiento longitudinal y transferencia del sistema a otros entornos, con menor fricción para la validación multicéntrica.

## **9. Síntesis y dirección futura.**

Comparado con el antecedente, este trabajo no se limitó a mejorar “métricas”, sino a mejorar decisiones clínicas: el balanceo por descarte (frente a oversampling), la calibración y la explicabilidad dual (Grad-CAM/LIME) ofrecieron un modelo más justo, fiable y auditable, especialmente en el eje crítico Normal–DCL destacado por el estudio de 2023. A futuro, la integración de la interfaz previa con superposiciones XAI y probabilidades calibradas, sumada a una validación con más jueces y más casos, permitirá capitalizar lo mejor de ambos enfoques y avanzar hacia una herramienta clínicamente lista, con transparencia y control de calidad de extremo a extremo.

## DISCUSION

El desarrollo del presente proyecto evidenció que la combinación de aprendizaje profundo con Inteligencia Artificial Explicable (XAI) no solo representó un avance técnico, sino también un paso fundamental hacia la adopción responsable y confiable de sistemas de IA en entornos clínicos. La problemática abordada la detección temprana del deterioro cognitivo mediante el análisis automatizado del Test de la Figura Compleja de Rey, exigió un enfoque que equilibrara precisión predictiva con interpretabilidad, dado que las decisiones derivadas de este tipo de herramientas impactan directamente en procesos diagnósticos y, por ende, en la calidad de vida de los pacientes.

Los resultados obtenidos confirmaron que la integración de XAI, a través de técnicas como Grad-CAM y LIME, permitió diagnosticar sesgos internos del modelo, tales como la dependencia excesiva en rasgos locales y la falta de atención a la estructura global del objeto. Estas evidencias no solo facilitaron la comprensión del comportamiento del modelo, sino que guiaron la implementación de estrategias correctivas fundamentadas, como la aumentación orientada a la forma, la regularización focalizada y el fine-tuning selectivo de capas profundas. Este ciclo iterativo, hallazgo, acción y verificación demostró que la explicabilidad no debe considerarse un complemento, sino un componente metodológico central en proyectos donde la transparencia y la trazabilidad son requisitos críticos.

Asimismo, la discusión sobre la importancia del proyecto trasciende el ámbito técnico. La propuesta respondió a una necesidad clínica real: optimizar la detección temprana del deterioro cognitivo en contextos donde los recursos humanos y el tiempo son limitados. Al ofrecer una herramienta capaz de automatizar la clasificación y, al mismo tiempo, justificar sus decisiones mediante explicaciones visuales y locales, se sentaron las bases para una solución que no solo mejora la eficiencia diagnóstica, sino que

también incrementa la confianza del personal de salud en la tecnología, reduciendo la resistencia a su adopción.

Por otra parte, la incorporación de XAI aportó un valor añadido en términos de auditoría y control de calidad. La posibilidad de visualizar qué regiones de la imagen influyeron en la predicción permitió detectar correlaciones espurias, validar la coherencia clínica del modelo y establecer umbrales de confianza para la toma de decisiones asistida. Este enfoque se alinea con las recomendaciones internacionales sobre IA confiable y ética, que enfatizan la necesidad de modelos interpretables, auditables y seguros en aplicaciones médicas.

El proyecto presentado representa un avance sustancial respecto al trabajo desarrollado por Cruz (2023), no solo en términos de arquitectura y métricas, sino en la metodología orientada a la explicabilidad y la trazabilidad clínica. Mientras el estudio anterior demostró la viabilidad técnica de un clasificador basado en CNN y la utilidad operativa de una interfaz para reducir tiempos de análisis, la presente investigación amplió el alcance hacia la interpretabilidad del modelo, incorporando técnicas de Inteligencia Artificial Explicable (XAI) como Grad-CAM y LIME, junto con calibración probabilística. Este cambio metodológico no se limitó a añadir explicaciones visuales, sino que transformó la forma en que se evalúa y ajusta el modelo, permitiendo detectar sesgos, orientar estrategias de regularización y validar la coherencia espacial de las predicciones.

Desde la perspectiva de adopción clínica, este avance influye directamente en la fase de integración con el personal de salud. El costo de lectura e interpretación por parte de profesionales neuropsicológicos es un factor crítico: aunque la interfaz previa redujo el tiempo de análisis de cada prueba de varios minutos a segundos, la ausencia de explicaciones limitaba la confianza del clínico en la herramienta. En este

proyecto, la generación de mapas Grad-CAM y explicaciones locales mediante LIME introduce un nuevo desafío: la interpretación de estas visualizaciones por personal no especializado en ingeniería. Si bien estas técnicas aportan transparencia, su comprensión requiere un puente semántico entre la representación computacional y el lenguaje clínico. Por ello, se diseñaron salidas gráficas intuitivas y reportes complementarios que facilitan la lectura por parte del profesional, reduciendo la curva de aprendizaje y el riesgo de rechazo tecnológico.

Este enfoque metodológico contribuye a la madurez del sistema para fases de adopción real, ya que la explicabilidad no solo mejora la confianza, sino que habilita mecanismos de auditoría y control de calidad exigidos por normativas éticas y regulatorias. Además, la posibilidad de justificar cada predicción con evidencia visual y probabilidades calibradas permite establecer umbrales de decisión y protocolos de derivación, integrando la herramienta en flujos clínicos sin sustituir el juicio profesional. En síntesis, el avance logrado no se limita a optimizar métricas, sino que redefine la interacción entre la IA y el personal médico, transformando la herramienta en un sistema interpretable, auditable y clínicamente viable, condición indispensable para su implementación en entornos sanitarios.

## SUGERENCIAS

A partir de los resultados obtenidos y de las conclusiones previamente expuestas, se recomendó continuar el proyecto siguiendo una ruta que consolidara la robustez técnica del modelo y su aplicabilidad clínica. En primer lugar, se sugirió ampliar y diversificar el conjunto de datos, incorporando imágenes provenientes de diferentes instituciones y contextos, con el fin de reducir sesgos y mejorar la capacidad de generalización del sistema. Esta ampliación debería complementarse con estrategias de aumentación guiadas por XAI, aplicando técnicas como Cutout o GridMask en regiones donde las explicaciones Grad-CAM indicaron una atención excesiva, promoviendo así un aprendizaje más equilibrado.

En segundo término, se propuso fortalecer la validación experimental mediante la implementación de validación cruzada estratificada (K-fold), lo que permitiría obtener métricas más estables y representativas. Asimismo, se recomendó priorizar métricas macro, como el F1-macro, para evaluar el desempeño global del modelo en escenarios con desbalance de clases, evitando depender únicamente de la exactitud global.

Otra línea de mejora consistió en explorar arquitecturas híbridas que combinaran convoluciones con mecanismos de atención (p. ej., CBAM o SE blocks), así como evaluar Vision Transformers adaptados a bocetos, comparando su rendimiento con el de CNNs tradicionales. Estas pruebas deberían realizarse bajo un marco metodológico que mantenga la interpretabilidad como requisito central, asegurando la compatibilidad con técnicas XAI.

En cuanto a la integración de explicabilidad, se recomendó profundizar el uso de XAI como herramienta iterativa, incorporando paneles comparativos sistemáticos que documenten la redistribución de la

atención antes y después de cada ajuste. Además, se sugirió complementar las explicaciones visuales con reportes textuales que indiquen qué rasgos influyeron en la decisión, facilitando la comprensión por parte del personal clínico.

Finalmente, se propuso incorporar un proceso avanzado de segmentación en la metodología, orientado a aislar puntos cruciales del TFCR como intersecciones, diagonales, figuras internas y elementos periféricos que son determinantes en la calificación neuropsicológica. Esta segmentación permitiría dividir la imagen en regiones semánticamente significativas, facilitando la correlación entre la presencia, ubicación y proporción de estos componentes y el diagnóstico clínico. Al trabajar sobre segmentos definidos, el modelo podría aprender relaciones estructurales más finas, reduciendo la dependencia de patrones globales y mejorando la sensibilidad ante errores visoconstructivos sutiles, característicos del deterioro cognitivo leve.

Además, esta estrategia potenciaría el uso de XAI, ya que las explicaciones generadas por Grad-CAM y LIME podrían enfocarse en áreas específicas del dibujo, ofreciendo mapas de atención más interpretables y alineados con criterios clínicos. Por ejemplo, un mapa que muestre atención en la cruz central o en el rectángulo principal tendría mayor valor diagnóstico que uno disperso sobre el fondo. Complementar estas visualizaciones con métricas de correlación entre segmentos y predicciones permitiría construir reportes cuantitativos y cualitativos, integrando evidencia visual y numérica para el profesional de la salud.

En términos de aplicabilidad, la segmentación no solo incrementaría la resolución funcional del modelo, sino que también facilitaría la adopción clínica, al presentar explicaciones más claras y centradas en elementos que el neuropsicólogo reconoce como relevantes. Esta mejora metodológica, combinada con

paneles comparativos y calibración probabilística, consolidaría un sistema robusto, interpretable y confiable, capaz de evolucionar hacia una herramienta validada en entornos reales.

## **Anexos**

**Anexo NO.1. Códigos Fase 1.**

**Anexo NO. 2. Segundo modelo de red neuronal. Fase 2.**

**Anexo NO. 3. Tercer modelo de red neuronal. Fase 3.**

**Anexo NO. 4. XAI.py**

**Anexo NO. 5. Código para la clasificación de imágenes por carpetas. Link de acceso a la carpeta en donde se encuentran los archivos, programas, base de datos y bancos de pruebas de la tesis.**

**<https://unbosqueeduco->**

**[my.sharepoint.com/:f:/g/personal/jbernalg\\_unbosque\\_edu\\_co/EgMLi3GPb1JKhMtXuB0-](https://unbosqueeduco-my.sharepoint.com/:f:/g/personal/jbernalg_unbosque_edu_co/EgMLi3GPb1JKhMtXuB0-)**

**[nqYBvaN2su\\_I2xO2UT6i52jQsQ?e=ZJHgB3](https://unbosqueeduco-my.sharepoint.com/:f:/g/personal/jbernalg_unbosque_edu_co/EgMLi3GPb1JKhMtXuB0-nqYBvaN2su_I2xO2UT6i52jQsQ?e=ZJHgB3)**

## BIBLIOGRAFÍA

- Ardila, A., Rosselli, M., & Puente, A. E. (2005). *Neuropsychological evaluation of the Spanish speaker*. Springer.
- Guerrero Barragán, A., Lucumí Cuesta, D. I., Gómez, I. E., & Lawlor, B. (2023). *Análisis situacional del deterioro cognitivo en Colombia* (No. 45). Escuela de Gobierno Alberto Lleras Camargo, Universidad de los Andes. <https://gobierno.uniandes.edu.co/wp-content/uploads/NP-45.pdf>
- Maldonado, L. V., Salazar, A. M., Puente, C., & Ávila, J. (2024). Sistema de apoyo diagnóstico de deterioro cognitivo basado en la figura compleja de Rey y redes neuronales.
- Pontón, M. O., Satz, P., Herrera, L., Ortiz, F., Urrutia, C. P., Young, R., D'Elia, L., & Namerow, N. (1996). Rey-Osterrieth Complex Figure Test: A normative study with a Spanish speaking pediatric sample. Neuropsychology Department, UCLA School of Medicine. [https://web.teaediciones.com/Ejemplos/RCFT\\_REY\\_web.pdf](https://web.teaediciones.com/Ejemplos/RCFT_REY_web.pdf)
- Rey, A. (1997). Rey: Test de copia y de reproducción de memoria de figuras geométricas complejas. Madrid: TEA ediciones.
- Gonzalez, R. C., & Woods, R. E. (2018). *Digital Image Processing* (4th ed.). Pearson.
- Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.
- Zhou, S. K., Greenspan, H., & Shen, D. (Eds.). (2019). *Deep learning for medical image analysis* (2nd ed.). Academic Press.

Petersen, R. C., Lopez, O., Armstrong, M. J., Getchius, T. S., Ganguli, M., Gloss, D., ... & Sager, M. (2018). Practice guideline update summary: Mild cognitive impairment. *Neurology*, 90(3), 126–135. <https://doi.org/10.1212/WNL.0000000000004826>

Ngandu, T., Lehtisalo, J., Solomon, A., Levälähti, E., Ahtiluoto, S., Antikainen, R., ... & Kivipelto, M. (2015). A 2-year multidomain intervention of diet, exercise, cognitive training, and vascular risk monitoring versus control to prevent cognitive decline in at-risk elderly people (FINGER): a randomised controlled trial. *The Lancet*, 385(9984), 2255–2263. [https://doi.org/10.1016/S0140-6736\(15\)60461-5](https://doi.org/10.1016/S0140-6736(15)60461-5)

Cruz Prieto, I. C. (2024). Sistema de detección temprana de deterioro cognitivo a través del análisis del test de figura compleja de Rey, haciendo uso de visión artificial [Trabajo de grado, Universidad El Bosque]. Repositorio Institucional Universidad El Bosque. <https://repositorio.unbosque.edu.co/bitstreams/af926399-9e7b-4f1b-831d-0fdf0628c9e7/download>

Gonzalez, R. C., & Woods, R. E. (2018). *Digital image processing* (4th ed.). Pearson.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Organización Mundial de la Salud. (2021). Demencia. <https://www.who.int/es/news-room/factsheets/detail/dementia>

Alzheimer's Disease International. (2022). *World Alzheimer Report 2022: Life after diagnosis: Navigating treatment, care and support*. <https://www.alzint.org/u/World-Alzheimer-Report-2022.pdf>

Howieson, D. (2019). Current limitations of neuropsychological tests and assessment procedures. *The Clinical Neuropsychologist*, 33(2), 200–208. <https://doi.org/10.1080/13854046.2018.1552762>

MathWorks. (2022). MATLAB documentation. <https://www.mathworks.com/help/matlab/>

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). Handbook of normative data for neuropsychological assessment (2nd ed.). Oxford University Press.

Programa Iberoamericano de Cooperación sobre la Situación de las Personas Adultas Mayores en la Región. (2023). Deterioro cognitivo en los adultos mayores y la importancia de su detección temprana. Iberoamérica Mayores. <https://iberoamericamayores.org/wp-content/uploads/2023/12/03-deteriorocognitivo.pdf>

Rey, A. (1941). L'examen psychologique dans les cas d'encéphalopathie traumatique. Archives de Psychologie, 28, 215–285.

Google (2025). Fotos de imágenes de números. Recuperado de [https://www.google.com/search?sca\\_esv=d915882ffb3cd600&sxsrf=AE3TifPIApziMRhL8P1gZjLNcguyzBN\\_jw:1761237992249&udm=2&fbs=AIjPjHx4nJjfGojPVHhEACUHPiMQht6\\_BFq6vBIoFFRK7qchKG1cRgcE0P7z3SNizmlu0QnAo31FcWSm9PsQnW8mPH21BwyyvooJOPmu9LTn8mR6PIOJzAXroCq8YBrQpcoigyipB\\_2YjPLKd9ZowU3VbhEiBe3-7YX1RDnQwfYXfM6Pjf2zrPpCPeTRd5BAV8SF4xRB7InzU9u4Sx45XerBzSOQ5Y3FBb4umD2P77KDD1eeRiEo7wk&q=Imagenes+de+numeros&sa=X&sqi=2&ved=2ahUKEwiWjZLn4rqQAxUnTDABHZqpH5QQtKgLegQIFhAB&biw=1536&bih=729&dpr=1.25](https://www.google.com/search?sca_esv=d915882ffb3cd600&sxsrf=AE3TifPIApziMRhL8P1gZjLNcguyzBN_jw:1761237992249&udm=2&fbs=AIjPjHx4nJjfGojPVHhEACUHPiMQht6_BFq6vBIoFFRK7qchKG1cRgcE0P7z3SNizmlu0QnAo31FcWSm9PsQnW8mPH21BwyyvooJOPmu9LTn8mR6PIOJzAXroCq8YBrQpcoigyipB_2YjPLKd9ZowU3VbhEiBe3-7YX1RDnQwfYXfM6Pjf2zrPpCPeTRd5BAV8SF4xRB7InzU9u4Sx45XerBzSOQ5Y3FBb4umD2P77KDD1eeRiEo7wk&q=Imagenes+de+numeros&sa=X&sqi=2&ved=2ahUKEwiWjZLn4rqQAxUnTDABHZqpH5QQtKgLegQIFhAB&biw=1536&bih=729&dpr=1.25)

Petersen, R. C., Lopez, O., Armstrong, M. J., Getchius, T. S., Ganguli, M., Gloss, D., ... & RaeGrant, A. (2014). Practice guideline update summary: Mild cognitive impairment. Neurology, 83(10), 980–986. <https://doi.org/10.1212/WNL.0000000000000797>

Ruiz, S. (2000). Evaluación neuropsicológica en demencias. Revista Colombiana de Psiquiatría, 29(2), 117–132. <https://www.scielo.org.co/pdf/rcp/v29n2/v29n2a09.pdf>

Sammut, C., & Webb, G. I. (Eds.). (2017). Encyclopedia of machine learning and data mining. Springer.

- Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 60. <https://doi.org/10.1186/s40537-019-0197-0>
- Van Rossum, G., & Drake, F. L. (2009). Python 3 reference manual. CreateSpace.
- Weiss, L. G., & Saklofske, D. H. (2020). WISC-V assessment and interpretation: Scientist–practitioner perspectives (2nd ed.). Academic Press.
- Zhou, S. K., Greenspan, H., & Shen, D. (Eds.). (2019). Deep learning for medical image analysis (2nd ed.). Academic Press.
- Park, J. Y., Seo, E. H., Yoon, H.-J., Won, S., & Lee, K. H. (2023). Automating Rey Complex Figure Test scoring using a deep learning-based approach: a potential large-scale screening tool for cognitive decline. *Alzheimer's Research & Therapy*, 15(1), 145. <https://doi.org/10.1186/s13195-023-01283-w>
- Langer, N., Weber, M., Vieira, B. H., Strzelczyk, D., Wolf, L., Pedroni, A., ... & Zhang, C. (2024). A deep learning approach for automated scoring of the Rey-Osterrieth complex figure. *eLife*, 13, RP96017. <https://doi.org/10.7554/eLife.96017>
- Guerrero-Martín, J., Díaz-Mardomingo, M. C., García-Herranz, S., Martínez-Tomás, R., & Rincón, M. (2024). A benchmark for Rey-Osterrieth complex figure test automatic scoring. *Heliyon*, 10(23), e39883. <https://doi.org/10.1016/j.heliyon.2024.e39883>
- Vogt, J., Kloosterman, H., Vermeent, S., Van Elswijk, G., Dotsch, R., & Schmand, B. (2019). Automated scoring of the Rey-Osterrieth Complex Figure Test using a deep-learning algorithm. *Archives of Clinical Neuropsychology*, 34(6), 836–836. <https://doi.org/10.1093/arclin/acz035.04>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A., Ciompi, F., Ghafoorian, M., ... & van der Laak, J. A. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42, 60–88.

<https://doi.org/10.1016/j.media.2017.07.005>

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 60. <https://doi.org/10.1186/s40537-019-0197-0>

Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.

<https://doi.org/10.1038/nature14539>

Gonzalez, R. C., & Woods, R. E. (2018). *Digital image processing* (4th ed.). Pearson.

Hawkins, D. M. (2004). *The problem of overfitting*. *Journal of Chemical Information and Computer Sciences*, 44(1), 1-12. <https://doi.org/10.1021/ci0342472>

Prechelt, L. (1998). *Automatic early stopping using cross validation: Quantifying the criteria*. *Neural Networks*, 11(4), 761–767. [https://doi.org/10.1016/S0893-6080\(98\)00010-0](https://doi.org/10.1016/S0893-6080(98)00010-0)

Shorten, C., & Khoshgoftaar, T. M. (2019). *A survey on Image Data Augmentation for Deep Learning*. *Journal of Big Data*, 6(60), 1-48. <https://doi.org/10.1186/s40537-019-0197-0>

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). *Dropout: A simple way to prevent neural networks from overfitting*. *Journal of Machine Learning Research*, 15(56), 1929–1958. <https://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf>

Wang, P., Liu, L., Shen, C., Huang, Z., van den Hengel, A., & Shen, H. T. (2020). *Image similarity using deep CNN and curriculum learning*. ACM Transactions on Multimedia Computing, Communications, and Applications, 16(1), 1-21. <https://doi.org/10.1145/3362123>

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>

Chen, P., Liu, S., Zhao, H., Wang, X., & Jia, J. (2024). GridMask data augmentation. arXiv preprint arXiv:2001.04086 (v3). <https://arxiv.org/abs/2001.04086>

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. <https://arxiv.org/abs/1702.08608>

European Commission. High-Level Expert Group on AI. (2019, April 8). Ethics guidelines for trustworthy AI. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning. <https://arxiv.org/abs/1706.04599>

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. <https://arxiv.org/abs/1602.04938>

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618–626. [https://openaccess.thecvf.com/content\\_ICCV\\_2017/papers/Selvaraju\\_Grad-CAM\\_Visual\\_Explanations\\_ICCV\\_2017\\_paper.pdf](https://openaccess.thecvf.com/content_ICCV_2017/papers/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.pdf)

Shanmugam, D., Blalock, D., Balakrishnan, G., & Gutttag, J. (2021). Better aggregation in test-time augmentation. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

[https://openaccess.thecvf.com/content/ICCV2021/papers/Shanmugam\\_Better\\_Aggregation\\_in\\_Test-Time\\_Augmentation\\_ICCV\\_2021\\_paper.pdf](https://openaccess.thecvf.com/content/ICCV2021/papers/Shanmugam_Better_Aggregation_in_Test-Time_Augmentation_ICCV_2021_paper.pdf)

Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2020). Random erasing data augmentation. Proceedings of the AAAI Conference on Artificial Intelligence, 34(7), 13001–13008.

<https://doi.org/10.1609/aaai.v34i07.7000>

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(11), 2274–2282.

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the Robustness of Interpretability Methods. arXiv preprint arXiv:1806.08049.

DeVries, T., & Taylor, G. W. (2017). Improved regularization of convolutional neural networks with Cutout. arXiv preprint arXiv:1708.04552.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7132–7141.

Kornblith, S., Shlens, J., & Le, Q. V. (2019). Do better ImageNet models transfer better? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2661–2671.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2980–2988.

- Loshchilov, I., & Hutter, F. (2017). SGDR: Stochastic Gradient Descent with Warm Restarts. In International Conference on Learning Representations (ICLR).
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 1135–1144.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618–626.
- Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1), 60.
- Simard, P. Y., Steinkraus, D., & Platt, J. C. (2003). Best practices for convolutional neural networks applied to visual document analysis. In Proceedings of the 7th International Conference on Document Analysis and Recognition, 958–963.
- Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), 3–19.
- Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, 27.
- Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 6023–6032.
- Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). mixup: Beyond Empirical Risk Minimization. In International Conference on Learning Representations (ICLR).
- Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2020). Random Erasing Data Augmentation.

AAAI Conference on Artificial Intelligence, 34(07), 13001–13008.