Pontificia Universidad
JAVERIANA
Cali

Facultad de Ingeniería y Ciencias

Ingeniería Biomédica

Informe Final de Trabajo de Grado

Desarrollo de un sistema predictivo para el apoyo en el diagnóstico temprano de la enfermedad de Alzheimer mediante inteligencia artificial y estudios PET

Alejandra Bolaños Aldana
Nicoll Dayana Castillo Estacio

Director
Dr. Hernán Darío Vargas Cardona

9 de febrero de 2026

Santiago de Cali, 9 de febrero de 2026

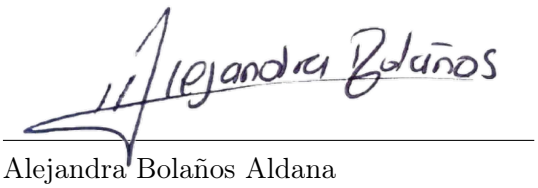
Señores
Pontificia Universidad Javeriana – Cali
Dr. Hernán Camilo Rocha Niño
Decano
Facultad de Ingeniería y Ciencias
Ciudad

Cordial Saludo.

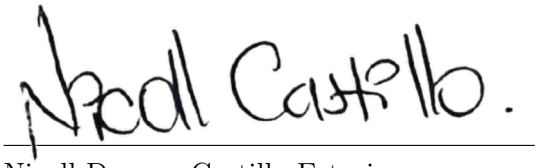
Por medio de la presente nos permitimos presentarle el Trabajo de Grado titulado "Desarrollo de un sistema predictivo para el apoyo en el diagnóstico temprano de la enfermedad de Alzheimer mediante inteligencia artificial y estudios PET".

Esperamos que este trabajo reúna todos los requisitos académicos, cumpla el propósito para el cual fue creado y sirva de apoyo para futuros proyectos relacionados con la profesión.

Atentamente,



Alejandra Bolaños Aldana



Nicoll Dayana Castillo Estacio

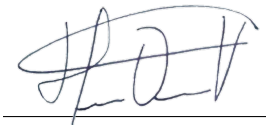
Santiago de Cali, 9 de febrero de 2026

Señores
Pontificia Universidad Javeriana – Cali
Dr. Hernán Camilo Rocha Niño
Decano
Facultad de Ingeniería y Ciencias
Ciudad

Cordial Saludo.

Certifico que el presente Trabajo de Grado titulado "Desarrollo de un sistema predictivo para el apoyo en el diagnóstico temprano de la enfermedad de Alzheimer mediante inteligencia artificial y estudios PET", realizado por Alejandra Bolaños Aldana y Nicoll Dayana Castillo Estacio, estudiantes de Ingeniería Biomédica, se encuentra terminado y puede ser presentado para su sustentación.

Atentamente,



Dr. Hernán Darío Vargas Cardona
Director Trabajo de Grado

Agradecimientos

Agradecemos a Dios por guiarnos y fortalecernos a lo largo de este proceso, brindándonos la sabiduría, la perseverancia y la fe necesarias para superar cada desafío y culminar con éxito esta etapa fundamental de nuestras vidas.

A nuestros padres y hermanos, por su amor incondicional, apoyo constante y confianza en nuestros sueños. Su comprensión, palabras de aliento y acompañamiento fueron pilares esenciales durante todo este camino. Asimismo, agradecemos a nuestros familiares por su cariño, oraciones y buenos deseos, que nos acompañaron durante toda esta travesía.

A nuestros amigos y a todas las personas que contribuyeron, directa o indirectamente, al desarrollo de este trabajo, gracias por su acompañamiento, apoyo y valiosos aportes a lo largo de este proceso.

A nuestro director, el Dr. Hernán Darío Vargas Cardona, le agradecemos sinceramente por su orientación, dedicación y conocimientos compartidos, los cuales fueron esenciales para el desarrollo riguroso de este proyecto. Su guía experta, paciencia y valiosas sugerencias contribuyeron de manera significativa al logro de los objetivos propuestos.

A nosotras mismas, por la perseverancia, el compromiso y la dedicación demostrados a lo largo de la carrera, incluso en los momentos de mayor dificultad. Este trabajo representa el esfuerzo conjunto que nos permitió superar los desafíos del proceso y alcanzar la meta de convertirnos en ingenieras.

A la Pontificia Universidad Javeriana Cali, a la Facultad de Ingeniería y Ciencias y a sus docentes, por la formación integral y las herramientas académicas brindadas durante este proceso.

Agradecimientos especiales de Alejandra

Agradezco de manera especial a mi mejor amiga por su constante presencia, escucha y paciencia a lo largo de este proceso. Gracias por acompañarme en los momentos de cansancio y dificultad, y por brindarme siempre palabras de ánimo que me impulsaron a continuar.

Agradecimientos especiales de Nicoll

Agradezco profundamente a Juan y a Giselly por su amor, apoyo incondicional y confianza en mí y en mis capacidades. Sus palabras de aliento, comprensión y acompañamiento fueron fundamentales para mantener la motivación y la fortaleza necesarias durante este proceso.

Finalmente, este logro se dedica a los pacientes y a las familias afectadas por la enfermedad de Alzheimer, con la esperanza de que investigaciones como esta contribuyan al diagnóstico temprano y a la mejora de su calidad de vida.

Glosario

Símbolos

δ_i	Tiempo límite de ejecución asociado a la i-ésima tarea, expresado en milisegundos [ms].
τ_i	i-ésima tarea considerada dentro de un sistema o proceso computacional.
k	Número de vecinos considerados en el algoritmo *K-Nearest Neighbors*.
μ	Media estadística estimada para una variable o distribución.
σ	Desviación estándar de una distribución de probabilidad.
σ^2	Varianza estadística asociada a una distribución de probabilidad.
$[0, 1]$	Intervalo utilizado para la normalización de intensidades de imagen.
$91 \times 109 \times 91$	Dimensiones espaciales estándar de los volúmenes PET preprocesados en vóxeles.
$p1, p99$	Percentiles 1 y 99 utilizados en la normalización robusta de intensidades de imagen.
X, Y, Z	Dimensiones espaciales en los ejes axial, coronal y sagital respectivamente, medidas en milímetros [mm].
N	Número total de muestras o estudios en un conjunto de datos.
^{18}F	Flúor-18, isótopo radiactivo utilizado en radiofármacos PET.
^{11}C	Carbono-11, isótopo radiactivo utilizado en radiofármacos PET.
TP	Verdaderos Positivos (*True Positives*).
TN	Verdaderos Negativos (*True Negatives*).
FP	Falsos Positivos (*False Positives*).
FN	Falsos Negativos (*False Negatives*).
mm	Milímetros, unidad de medida para dimensiones de vóxel.
ms	Milisegundos, unidad de medida temporal.

Acrónimos y Abreviaturas

AD	*Alzheimer's Disease* (Enfermedad de Alzheimer).
ADNI	*Alzheimer's Disease Neuroimaging Initiative*.
API	*Application Programming Interface*.
AUC	*Area Under the Curve* (Área bajo la curva).
CN	*Cognitively Normal* (Cognitivamente normal).
CNN	*Convolutional Neural Network* (Red neuronal convolucional).

DICOM	*Digital Imaging and Communications in Medicine.*
DL	*Deep Learning* (Aprendizaje profundo).
EA	Enfermedad de Alzheimer.
FDG	Fluorodeoxiglucosa ([^{18}F]FDG).
IEEE	*The Institute of Electrical and Electronic Engineers.*
IoT	*Internet of Things.*
JSON	*JavaScript Object Notation.*
KNN	*K-Nearest Neighbors.*
MCI	*Mild Cognitive Impairment* (Deterioro cognitivo leve).
ML	*Machine Learning* (Aprendizaje automático).
MMSE	*Mini-Mental State Examination.*
MRI	*Magnetic Resonance Imaging* (Resonancia magnética).
NIfTI	*Neuroimaging Informatics Technology Initiative.*
PET	*Positron Emission Tomography* (Tomografía por emisión de positrones).
PIB	*Pittsburgh Compound B* (^{11}C-PIB).
ResNet	*Residual Network* (Red residual).
ROC	*Receiver Operating Characteristic.*
SMOTE	*Synthetic Minority Over-sampling Technique.*

Términos

Aprendizaje profundo	Subcampo del aprendizaje automático basado en redes neuronales profundas capaces de aprender representaciones jerárquicas a partir de grandes volúmenes de datos.
Arquitectura residual	Diseño de red neuronal que incorpora conexiones de salto para facilitar el entrenamiento de redes profundas y mitigar el desvanecimiento del gradiente.
Batch Normalization	Técnica utilizada para normalizar las activaciones intermedias de una red neuronal con el fin de estabilizar y acelerar el entrenamiento.
Callback	Mecanismo que permite ejecutar acciones específicas durante el entrenamiento de un modelo, como detención anticipada o ajuste dinámico de la tasa de aprendizaje.
Clasificación multiclase	Problema de aprendizaje automático en el que una muestra puede pertenecer a una de tres o más clases posibles.
Curva ROC	Representación gráfica del compromiso entre sensibilidad y especificidad de un clasificador para diferentes umbrales de decisión.
Desbalanceo de clases	Condición en un conjunto de datos donde algunas categorías tienen significativamente más muestras que otras, afectando potencialmente el desempeño del modelo.

Dropout	Técnica de regularización que desactiva aleatoriamente neuronas durante el entrenamiento para prevenir el sobreajuste.
Early stopping	Estrategia de regularización que detiene el entrenamiento cuando el desempeño en validación deja de mejorar, previniendo el sobreajuste.
Enmascaramiento cerebral	Proceso de eliminación de tejidos extracerebrales de las imágenes médicas mediante la aplicación de una máscara binaria.
Especificidad	Proporción de casos negativos correctamente identificados por un clasificador.
F1-Score	Métrica que combina precisión y recall mediante su media armónica, útil en problemas con desbalanceo de clases.
Generalización	Capacidad de un modelo de mantener su desempeño en datos no vistos durante el entrenamiento.
Heatmap	Representación gráfica matricial que utiliza colores para visualizar la intensidad o frecuencia de valores, comúnmente empleada para matrices de confusión.
Hipometabolismo	Reducción del metabolismo energético en regiones cerebrales específicas, observable mediante FDG-PET.
Información mutua	Medida estadística de dependencia entre dos variables, utilizada como métrica de similitud en algoritmos de registro de imágenes.
Interpolación	Proceso matemático utilizado para estimar valores intermedios durante el redimensionamiento de imágenes.
Joblib	Librería de Python optimizada para la serialización eficiente de objetos que contienen grandes arreglos numéricos.
Matriz de confusión	Tabla que resume el desempeño de un clasificador mostrando predicciones correctas e incorrectas por clase.
Neuroimagen	Conjunto de técnicas que permiten visualizar la estructura y función del sistema nervioso.
Normalización de intensidades	Proceso de escalado de valores de imagen para garantizar consistencia entre estudios adquiridos bajo diferentes condiciones.
Ovillos neurofibrilares	Estructuras patológicas formadas por proteínas Tau hiperfosforiladas, características de la enfermedad de Alzheimer.
Pickle	Módulo de Python que permite serializar y deserializar objetos complejos en formato binario.
Pipeline	Secuencia estructurada de etapas de preprocesamiento, entrenamiento y evaluación aplicada a los datos.
Placas amiloides	Depósitos extracelulares de proteína beta-amiloide en el tejido cerebral, marcador temprano de la enfermedad de Alzheimer.
Precisión	Proporción de predicciones positivas que son correctas, calculada como verdaderos positivos sobre total de positivos predichos.

Radiofármaco	Compuesto químico marcado con un isótopo radiactivo utilizado en estudios de medicina nuclear.
Recall	Proporción de casos positivos correctamente identificados, también conocida como sensibilidad.
Registro rígido	Proceso de alineación espacial de imágenes mediante transformaciones que preservan distancias y ángulos (traslación y rotación).
Sensibilidad	Capacidad de un clasificador para identificar correctamente los casos positivos, equivalente al recall.
Sobreajuste	Fenómeno en el que un modelo aprende patrones específicos del conjunto de entrenamiento que no generalizan a datos nuevos.
Transfer Learning	Estrategia de aprendizaje que reutiliza modelos preentrenados para resolver tareas relacionadas con menor cantidad de datos.
Trazador	Sustancia marcada radiactivamente utilizada para visualizar procesos biológicos específicos mediante técnicas de imagen molecular.
Vóxel	Unidad volumétrica tridimensional equivalente a un píxel en imágenes 3D.
LaTeX	Lenguaje de composición tipográfica orientado a la creación de documentos científicos y técnicos de alta calidad.

Resumen

La enfermedad de Alzheimer constituye la principal causa de demencia a nivel mundial, afectando a más de 55 millones de personas. No obstante, su diagnóstico temprano continúa siendo un desafío clínico relevante, dado que los métodos convencionales suelen identificar la enfermedad en fases avanzadas, cuando las alternativas terapéuticas son limitadas. En este contexto, el presente trabajo desarrolló un sistema predictivo orientado al apoyo en el diagnóstico temprano de la enfermedad de Alzheimer mediante el uso de inteligencia artificial aplicada a estudios de tomografía por emisión de positrones (PET). El objetivo principal consistió en integrar técnicas de aprendizaje profundo y aprendizaje automático para la clasificación de pacientes en tres categorías diagnósticas: cognitivamente normal, deterioro cognitivo leve y enfermedad de Alzheimer.

La metodología se fundamentó en el uso de datos del repositorio público ADNI, incorporando un total de 5,673 imágenes PET adquiridas con diferentes radiofármacos, así como 4,617 registros clínicos que incluyen variables sociodemográficas, cognitivas y genéticas. Para el análisis de neuroimagen, se implementaron y compararon tres arquitecturas de redes neuronales convolucionales tridimensionales: ResNet3D, un enfoque de *transfer learning* basado en ResNet10-3D preentrenado y la arquitectura VoxCNN3D. De manera complementaria, se evaluaron modelos clásicos de aprendizaje automático aplicados a datos tabulares, específicamente K-Nearest Neighbors, Naive Bayes y Random Forest. Adicionalmente, se desarrolló un modelo híbrido que integró las representaciones profundas extraídas por ResNet10-3D con variables clínicas procesadas mediante Random Forest, con el fin de aprovechar información multimodal.

Los resultados evidenciaron que el modelo híbrido alcanzó el mejor desempeño global, logrando una exactitud del 77.12 % en el conjunto de prueba, superando de manera significativa a los modelos individuales. En particular, el sistema obtuvo una precisión del 100 % para la clase Alzheimer, un *recall* del 94.92 % para la clase de controles normales y métricas balanceadas para la categoría de deterioro cognitivo leve, lo que refleja una adecuada capacidad discriminativa. Como parte del desarrollo tecnológico, se implementó una interfaz gráfica funcional mediante Gradio, la cual permite la carga de estudios PET, la captura de información clínica, la visualización multiplanar de las neuroimágenes y la generación automática de reportes diagnósticos en formatos TXT y PDF.

En conclusión, este trabajo demuestra que la integración multimodal de neuroimagen funcional y datos clínicos, mediante técnicas avanzadas de inteligencia artificial, mejora de forma sustancial el desempeño diagnóstico frente a enfoques unimodales. El sistema propuesto se perfila como una herramienta prometedora de apoyo al diagnóstico clínico, con potencial para fortalecer la detección temprana y la estratificación de pacientes dentro del espectro del deterioro cognitivo.

Palabras Clave: Enfermedad de Alzheimer, tomografía por emisión de positrones, aprendizaje profundo, redes neuronales convolucionales, transfer learning, modelo híbrido, diagnóstico asistido por computadora, neuroimagen funcional, inteligencia artificial en medicina.

Abstract

Alzheimer's disease is the leading cause of dementia worldwide, affecting more than 55 million people. Nevertheless, early diagnosis remains a critical clinical challenge, as conventional methods often identify the disease at advanced stages, when therapeutic options are limited. In this context, this work developed a predictive system to support the early diagnosis of Alzheimer's disease using artificial intelligence applied to positron emission tomography (PET) studies. The main objective was to integrate deep learning and machine learning techniques to classify patients into three diagnostic categories: cognitively normal, mild cognitive impairment, and Alzheimer's disease.

The methodology was based on data from the public ADNI repository, incorporating a total of 5,673 PET images acquired with different radiotracers, along with 4,617 clinical records, including sociodemographic, cognitive, and genetic variables. For neuroimaging analysis, three three-dimensional convolutional neural network architectures were implemented and compared: ResNet3D, a transfer learning approach based on a pre-trained ResNet10-3D model, and VoxCNN3D. In parallel, classical machine learning models were evaluated on tabular data, namely K-Nearest Neighbors, Naive Bayes, and Random Forest. Additionally, a hybrid model was developed by integrating the deep representations extracted by ResNet10-3D with clinical variables processed through Random Forest, enabling the exploitation of multimodal information.

The results demonstrated that the hybrid model achieved the best overall performance, reaching an accuracy of 77.12 % on the test set and significantly outperforming the individual models. Specifically, the system achieved perfect precision (100 %) for the Alzheimer's disease class, a recall of 94.92 % for cognitively normal controls, and balanced metrics for the mild cognitive impairment category, reflecting a robust discriminative capability. As part of the technological development, a functional graphical user interface was implemented using Gradio, allowing PET study uploads, clinical data entry, multiplanar neuroimaging visualization, and the automatic generation of diagnostic reports in TXT and PDF formats.

In conclusion, this work demonstrates that the multimodal integration of functional neuroimaging and clinical information through advanced artificial intelligence techniques substantially improves diagnostic performance compared to unimodal approaches. The proposed system represents a promising clinical decision support tool, with the potential to enhance early detection and patient stratification across the cognitive impairment spectrum.

Keywords: Alzheimer's disease, positron emission tomography, deep learning, convolutional neural networks, transfer learning, hybrid model, computer-aided diagnosis, functional neuroimaging, artificial intelligence in medicine.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Durante los últimos años, la enfermedad de Alzheimer (EA) se ha consolidado como una de las principales causas de demencia a nivel mundial, afectando a más de 55 millones de personas y generando un impacto profundo tanto en los pacientes como en sus familiares y cuidadores [7]. Esta patología neurodegenerativa progresiva se caracteriza por la pérdida gradual de funciones cognitivas, como la memoria, el lenguaje y la orientación espacial, así como por alteraciones conductuales y emocionales que deterioran de manera significativa la calidad de vida del paciente[8]. Su incidencia ha incrementado de forma proporcional al envejecimiento global de la población, convirtiéndose en un reto prioritario para los sistemas de salud pública debido a los altos costos de atención, la dependencia prolongada y las limitadas alternativas terapéuticas disponibles [9].

Uno de los principales desafíos en torno a la EA es su diagnóstico temprano. La identificación de la enfermedad en sus etapas iniciales permite implementar estrategias terapéuticas más efectivas, retrasar la progresión del deterioro cognitivo y optimizar los cuidados del paciente[10]. Sin embargo, los métodos diagnósticos convencionales, que incluyen la anamnesis clínica, las pruebas neuropsicológicas y las técnicas de neuroimagen estructural, presentan limitaciones al momento de detectar los cambios cerebrales sutiles que anteceden los síntomas clínicos evidentes [11]. Además, la interpretación de estas pruebas depende en gran medida del criterio clínico del especialista, lo cual introduce un nivel de subjetividad que puede retrasar la detección o generar errores diagnósticos [12].

En este contexto, las técnicas de neuroimagen funcional han adquirido un papel fundamental en el diagnóstico diferencial de la EA. Particularmente, la tomografía por emisión de positrones (PET) ha demostrado una gran capacidad para identificar alteraciones metabólicas cerebrales y acumulación de proteínas anormales, como la beta-amiloide y la tau, antes de la aparición de síntomas clínicos [13]. A través del uso de diferentes radiofármacos, como el ^{18}F-FDG, se pueden observar patrones de hipometabolismo en regiones específicas del cerebro, los cuales actúan como biomarcadores confiables del avance de la enfermedad. No obstante, la interpretación de los estudios PET sigue siendo un proceso complejo que depende de la experiencia del profesional, lo que motiva el desarrollo de herramientas automatizadas que apoyen la lectura objetiva y reproducible de estas imágenes[14, 15].

En los últimos años, la inteligencia artificial (IA) ha emergido como una herramienta poderosa en el ámbito biomédico, especialmente en el análisis automatizado de imágenes médicas[16, 17]. Dentro de este campo, las redes neuronales convolucionales tridimensionales (3D-CNN) han mostrado un notable desempeño en la detección de patrones complejos presentes en imágenes PET, superando los métodos convencionales en precisión diagnóstica [18]. Estas redes permiten procesar directamente los volúmenes cerebrales completos, preservando la información espacial tridimensional, lo cual es crucial para identificar las alteraciones metabólicas y estructurales características de la EA.

Complementariamente, los enfoques de aprendizaje automático (*machine learning*, ML) han sido empleados para analizar variables clínicas, cognitivas, demográficas y genéticas que influyen en la progresión de la enfermedad [19]. La combinación de estas variables con los resultados derivados del análisis de neuroimágenes ha dado lugar a modelos híbridos más robustos, capaces de ofrecer predicciones personalizadas basadas en el perfil integral del paciente. Estudios recientes han demostrado que integrar información como la edad, el sexo, los años de educación, los antecedentes clínicos y el genotipo APOE puede mejorar significativamente la sensibilidad y especificidad de los modelos predictivos [20, 21].

A partir de este panorama, el presente proyecto de investigación desarrolla un sistema computacional predictivo para el apoyo en el diagnóstico temprano de la enfermedad de Alzheimer mediante el uso de algoritmos de inteligencia artificial entrenados con estudios PET y variables demográficas, clínicas y genéticas. El sistema propuesto integra técnicas de aprendizaje profundo a través de arquitecturas de redes neuronales convolucionales tridimensionales, junto con modelos clásicos de *machine learning*, con el fin de identificar los distintos niveles clínicos de la enfermedad: cognitivamente normal (CN), deterioro cognitivo leve (MCI) y Alzheimer (AD).

El objetivo general de este trabajo consiste en desarrollar un sistema para el apoyo diagnóstico de la enfermedad de Alzheimer mediante algoritmos de inteligencia artificial entrenados con estudios de tomografía por emisión de positrones. Para alcanzar este propósito, se establecieron cuatro objetivos específicos que guiaron el desarrollo del proyecto. En primer lugar, se gestionó una base de datos de estudios PET provenientes de pacientes con diagnóstico de deterioro cognitivo, enfermedad de Alzheimer y sujetos de control. Posteriormente, se implementaron algoritmos basados en redes neuronales convolucionales y máquinas de aprendizaje entrenados con estudios PET que permiten identificar los distintos niveles de la enfermedad. Estos algoritmos fueron validados mediante la evaluación con métricas estandarizadas para clasificación. Finalmente, se diseñó una interfaz de visualización que permite a un usuario procesar estudios PET de manera interactiva.

La metodología adoptada es de tipo experimental y computacional, basándose en datos secundarios, anónimos y de acceso público provenientes del repositorio Alzheimer's Disease Neuroimaging Initiative (ADNI). Este repositorio ofrece tanto imágenes PET con diferentes radiofármacos como datos clínicos, neuropsicológicos y genéticos, lo que permite desarrollar una base de datos multimodal y confiable. El proyecto contempla diversas fases metodológicas que abarcan desde el preprocesamiento de imágenes y datos clínicos hasta la implementación, el entrenamiento y la evaluación de múltiples arquitecturas de aprendizaje automático.

El preprocesamiento de imágenes PET constituyó una etapa fundamental para garantizar la calidad y homogeneidad de los datos. Las imágenes originales fueron normalizadas espacialmente mediante el registro a un atlas de referencia; se aplicaron técnicas de normalización de intensidades para estandarizar los rangos dinámicos y se implementó el enmascaramiento cerebral para aislar el tejido de interés. El conjunto resultante fue dividido estratificadamente en conjuntos de entrenamiento, validación y prueba, preservando la representatividad de las tres categorías diagnósticas. De manera análoga, los datos sociodemográficos y clínicos fueron sometidos a un proceso riguroso de preprocesamiento que incluyó la codificación de variables categóricas, la imputación de valores faltantes y la estandarización de características continuas.

La implementación de modelos abarcó dos aproximaciones complementarias. Por un lado, se desarrollaron tres arquitecturas de redes neuronales convolucionales 3D con diferentes características: ResNet3D, diseñada específicamente para este estudio e incorporando bloques residuales que facilitan el entrenamiento de redes profundas; un modelo basado en *Transfer Learning* con ResNet10-3D preentrenada en datos médicos, aplicando un esquema de ajuste en dos fases; y VoxCNN3D, una arquitectura ligera orientada a la eficiencia computacional. Por otro lado, se implementaron cuatro modelos de *machine learning* utilizando características clínicas y demográficas: K-Nearest Neighbors optimizado mediante validación cruzada, una variante de KNN con técnicas de balanceo sintético de clases (SMOTE), un clasificador Naive Bayes Gaussiano y Random Forest con SMOTE optimizado mediante búsqueda exhaustiva de hiperparámetros.

Las estrategias de entrenamiento y validación fueron diseñadas para garantizar evaluaciones rigurosas y comparables entre todos los modelos. Los modelos de aprendizaje profundo fueron entrenados utilizando optimizadores adaptativos, funciones de pérdida apropiadas para la clasificación multiclase y mecanismos de regularización para prevenir el sobreajuste. Los modelos de *machine learning* emplearon validación cruzada estratificada y búsqueda sistemática de hiperparámetros. Todos los modelos fueron evaluados mediante un conjunto consistente de métricas que incluyen exactitud, precisión por clase, sensibilidad, especificidad, F1-score y análisis de curvas ROC, permitiendo una caracterización completa de su comportamiento diagnóstico. Se implementaron rigurosas medidas para prevenir la fuga de información, garantizando que todas las transformaciones de preprocesamiento se ajustaran exclusivamente al conjunto de entrenamiento.

Los conceptos fundamentales que articulan este proyecto abarcan diversas áreas del conocimiento de manera integrada e interdisciplinaria. En el ámbito de la neuroimagen médica, se profundiza en los principios físicos de la tomografía PET y el funcionamiento de los diferentes tipos de trazadores empleados para visualizar amiloide, tau y metabolismo cerebral. En el campo de la inteligencia artificial, se abordan los fundamentos de las redes neuronales convolucionales, incluidas las arquitecturas residuales y las estrategias de *transfer learning* para aprovechar el conocimiento preentrenado. Los algoritmos de aprendizaje automático se analizan desde sus fundamentos probabilísticos y geométricos, considerando métodos de ensemble y técnicas para manejar el desbalanceo de clases. Finalmente, se examinan las métricas de evaluación apropiadas para problemas de clasificación médica multiclase.

Desde una perspectiva ética y metodológica, todo el trabajo se realizó sobre datos secundarios, completamente anónimos y de acceso público provenientes de la iniciativa ADNI, cumpliendo rigurosamente con las normas éticas internacionales para investigaciones sin intervención directa sobre seres humanos. El repositorio ADNI proporciona un conjunto estandarizado y ampliamente validado que combina imágenes PET, datos clínicos y evaluaciones neuropsicológicas mediante instrumentos como el Mini-Mental State Examination (MMSE), así como variables genéticas como el genotipo APOE, constituyendo la base ideal para el desarrollo y validación de modelos predictivos robustos en el contexto de la enfermedad de Alzheimer.

El presente documento se estructura de manera lógica y secuencial para facilitar la comprensión del trabajo realizado. El Capítulo 2 presenta el planteamiento del problema, evidenciando la necesidad de desarrollar herramientas automáticas de apoyo al diagnóstico temprano. El Capítulo

3 desarrolla la justificación del estudio desde perspectivas teóricas, prácticas y metodológicas. El Capítulo 4 enuncia formalmente los objetivos que guiaron el desarrollo del proyecto. El Capítulo 5 aborda el marco de referencia de manera comprensiva, incluyendo fundamentos sobre la enfermedad de Alzheimer, técnicas de diagnóstico por neuroimagen, descripción del repositorio ADNI, fundamentos de inteligencia artificial y revisión de trabajos relacionados. El Capítulo 6 describe exhaustivamente los materiales y métodos empleados, incluyendo la caracterización de las bases de datos, procedimientos de preprocesamiento, especificaciones de los modelos implementados, estrategias de entrenamiento y herramientas utilizadas. El Capítulo 7 presenta los resultados obtenidos de manera estructurada, comenzando con el análisis exploratorio del dataset y posteriormente el desempeño detallado de cada modelo mediante múltiples métricas y visualizaciones comparativas. El Capítulo 8 sintetiza las conclusiones del estudio, respondiendo a los objetivos planteados. Finalmente, el Capítulo 9 propone líneas de trabajo futuro que podrían fortalecer la aplicabilidad clínica y la robustez metodológica del sistema desarrollado.

Capítulo 2

Planteamiento del Problema

La enfermedad de Alzheimer (EA) es una de las principales causas de demencia en la población adulta, constituyendo entre el 60 % y el 70 % de todos los casos a nivel mundial, lo que representa un desafío significativo para la salud pública debido a su creciente prevalencia e impacto a nivel personal, familiar y socioeconómico asociado. Según la Organización Mundial de la Salud (OMS), más de 55 millones de personas viven con demencia, y se espera que este número aumente drásticamente en las próximas décadas, en parte por el envejecimiento global de la población [22]. Por otro lado, en América Latina y el Caribe, el impacto proyectado es aún más alarmante, ya que se estima que el número de personas con demencia se triplicará para 2050, lo que agravará los desafíos de los sistemas de salud [23]. Ahora bien, este tipo de demencia se trata de una enfermedad neurodegenerativa progresiva que inicialmente se manifiesta con pérdida de memoria y dificultad para realizar actividades cotidianas, donde, a medida que avanza la enfermedad, se acompaña de una profunda alteración en las funciones cognitivas, emocionales y motoras. De igual manera, uno de los principales problemas en el manejo de esta enfermedad es su diagnóstico, que suele realizarse en etapas avanzadas, cuando los síntomas cognitivos ya son evidentes y las opciones terapéuticas son limitadas; este diagnóstico tardío no sólo retrasa la intervención terapéutica, sino que también tiene un impacto negativo en la calidad de vida de los pacientes y sus familias [7]. Dentro de este contexto, el diagnóstico de la EA tradicionalmente se ha basado en una combinación de métodos clínicos y pruebas neuropsicológicas, donde estas evaluaciones incluyen entrevistas detalladas con los pacientes y sus familiares, enfocándose en la historia clínica y los síntomas de pérdida de memoria y habilidades cognitivas. Además, las pruebas neuropsicológicas evalúan la memoria, la capacidad de atención, la resolución de problemas y la orientación espacial [24].

De igual manera, se ha utilizado la neuroimagen, particularmente la resonancia magnética (RM), la cual se ha consolidado como una herramienta clave en el diagnóstico, ya que permite identificar los cambios estructurales en el cerebro, como la atrofia del hipocampo, que es una de las primeras áreas afectadas por la EA. Por otro lado, se destaca la tomografía por emisión de positrones (PET), que permite visualizar la acumulación de proteínas beta-amiloide y tau, las cuales son características neuropatológicas de la EA [25][26]. Ahora bien, centrándose en la sensibilidad diagnóstica empleando estas técnicas convencionales por neuroimagen, se obtiene un 80 % a 94 % para RM y 94 % para PET; asimismo, una especificidad del 60 % al 100 % para RM y del 73 % al 78 % para PET. No obstante, esto es aún menor en las etapas tempranas de la enfermedad, debido a la falta de manifestaciones clínicas claras en sus fases iniciales [27].

En cuanto a Colombia, la situación es preocupante, ya que el país enfrenta un envejecimiento progresivo de su población. Según datos del estudio Salud, Bienestar y Envejecimiento (SABE) Colombia, realizado en Bogotá, alrededor del 10 % de los adultos mayores presentaron algún tipo

de deterioro cognitivo, con una prevalencia estimada de la enfermedad de Alzheimer del 1,8 % en la población general [22].

Además, según un estudio sobre la epidemiología y la carga de la EA en Colombia, se esperaba que para el año 2020 la población colombiana mayor de 69 años fuera de 6.4 millones, con al menos 263 casos de demencia. No obstante, actualmente aún se evidencia la carencia de datos epidemiológicos actualizados respecto al número de personas mayores que padecen la EA en Colombia, así como la carga económica que esta enfermedad impone en el sistema de salud colombiano, lo que dificulta la planificación adecuada de recursos y estrategias de manejo a largo plazo [28].

Por otra parte, investigaciones recientes han demostrado que los estudios PET son una de las técnicas de neuroimagen más prometedoras para la detección temprana de la EA, puesto que permiten identificar cambios funcionales y metabólicos asociados con la enfermedad incluso antes de la manifestación de síntomas clínicos. En particular, el uso de trazadores específicos para la detección de beta-amiloide y proteína tau ha permitido una caracterización más precisa de los procesos neuropatológicos de la EA en sus etapas iniciales. [29]

No obstante, la interpretación de los estudios PET sigue representando un desafío en la práctica clínica, ya que depende en gran medida de la experticia del especialista y puede estar sujeta a variabilidad en la lectura de los resultados. En donde esta falta de precisión en la interpretación de las imágenes dificulta su uso como herramienta de diagnóstico temprano en contextos clínicos rutinarios. Por lo que, ante esta problemática, el desarrollo de modelos predictivos basados en inteligencia artificial, específicamente con redes neuronales convolucionales (CNN), ha surgido como una alternativa para mejorar el análisis de estudios PET en la detección temprana de la EA, demostrando así la capacidad de identificar patrones complejos en las imágenes con una precisión superior a los métodos convencionales, alcanzando tasas de precisión del 80 % al 95 % [30, 17, 31]. Además, se ha evidenciado que la integración de enfoques de machine learning en conjunto con características demográficas, clínicas, cognitivas y genéticas puede aumentar significativamente la sensibilidad y especificidad del diagnóstico. Esta combinación permite construir modelos más robustos y personalizados, capaces de predecir el riesgo o la progresión de la enfermedad con mayor exactitud, permitiendo así mejores tomas de decisiones clínicas y la planificación de intervenciones tempranas [17, 32]. Entre estos enfoques, se ha explorado el uso de algoritmos como random forests, máquinas de vectores de soporte (SVM) y K-Nearest Neighbors (KNN), que han demostrado un buen desempeño en la clasificación de pacientes con diferentes grados de deterioro cognitivo, permitiendo integrar de manera efectiva datos heterogéneos provenientes de distintas fuentes clínicas y de neuroimagen. [33]

Por lo tanto, surge la necesidad de promover cómo se podría integrar de manera eficaz los datos que resultan de los estudios PET y de diversas características clínicas, demográficas, entre otras, para lograr un diagnóstico eficaz de la enfermedad de Alzheimer. En donde, resolver este problema no solo mejoraría la precisión diagnóstica, sino que también permitiría intervenciones terapéuticas más oportunas, mejorando así la calidad de vida de los pacientes y asegurando un enfoque más completo y efectivo en el diagnóstico y manejo de estas enfermedades neurodegenerativas [34].

Capítulo 3

Justificación

La enfermedad de Alzheimer (EA) representa una de las principales causas de demencia a nivel mundial, y su impacto continúa en aumento debido al envejecimiento progresivo de la población. En este contexto, la precisión en el diagnóstico se convierte en un factor determinante para la implementación de tratamientos oportunos y personalizados, los cuales pueden ralentizar la progresión de la enfermedad y mejorar la calidad de vida de los pacientes. Sin embargo, los métodos diagnósticos convencionales presentan limitaciones considerables, ya que suelen detectar la enfermedad en etapas avanzadas, cuando el deterioro cognitivo ya es significativo y las opciones terapéuticas se reducen considerablemente [35].

Ante esta problemática, el desarrollo de un modelo predictivo basado en inteligencia artificial (IA) que aproveche el análisis automatizado de imágenes PET surge como una alternativa prometedora. Estas imágenes permiten detectar de forma no invasiva alteraciones funcionales y moleculares en el cerebro mediante el uso de distintos radiofármacos, los cuales posibilitan visualizar patrones asociados a la acumulación de amiloide, tau y alteraciones en el metabolismo de la glucosa. La integración de estas trazas moleculares a través de herramientas de aprendizaje profundo (deep learning) puede facilitar la detección temprana de la enfermedad, incluso antes de la aparición de síntomas clínicos evidentes, contribuyendo así a mejorar la precisión diagnóstica y la capacidad de predicción del curso clínico de la EA [36].

Desde el punto de vista teórico, este proyecto contribuye al fortalecimiento del campo de la bioingeniería al proponer la combinación de modelos de aprendizaje profundo y aprendizaje automático en el procesamiento de imágenes médicas. La investigación plantea un enfoque interdisciplinar que integra conocimientos de neurociencia, ingeniería biomédica e inteligencia artificial, lo que permite avanzar en el desarrollo de herramientas tecnológicas con potencial de aplicación clínica. De esta manera, el trabajo no solo aporta a la literatura científica sobre diagnóstico asistido por IA, sino que también abre nuevas líneas de investigación en la detección temprana de enfermedades neurodegenerativas mediante biomarcadores de imagen.

En términos prácticos, la implementación de este modelo beneficiará a distintos actores dentro del sistema de salud. En primer lugar, los pacientes en riesgo de desarrollar Alzheimer podrán acceder a diagnósticos más tempranos y precisos, lo que permitirá la aplicación de estrategias terapéuticas en fases iniciales, aumentando la efectividad de los tratamientos y mejorando su calidad de vida. En segundo lugar, los familiares de los pacientes se verán favorecidos, ya que un diagnóstico oportuno facilitará la planificación del cuidado y la asistencia de sus seres queridos, reduciendo la carga emocional y económica asociada al manejo de la enfermedad en etapas avanzadas. Asimismo, el uso de herramientas computacionales para el análisis de neuroimágenes permitirá reducir la subjetividad en la interpretación de los resultados, ofreciendo al personal médico apoyo en la toma de decisiones

clínicas basadas en evidencia cuantitativa y reproducible [37].

Desde la perspectiva metodológica, el desarrollo de un sistema predictivo de este tipo implica la gestión y análisis de bases de datos médicas estandarizadas, como el repositorio público ADNI, y la implementación de arquitecturas tridimensionales de redes neuronales convolucionales (3D-CNN). A diferencia del diagnóstico convencional basado en cortes bidimensionales, el uso de datos volumétricos y modelos tridimensionales permite capturar la estructura completa del cerebro, identificando alteraciones metabólicas y anatómicas con mayor precisión [38]. Esto garantiza un enfoque analítico más robusto y adaptado a la complejidad de los datos biomédicos, favoreciendo resultados más consistentes y generalizables.

En síntesis, el desarrollo de un modelo predictivo basado en inteligencia artificial para el diagnóstico temprano de la enfermedad de Alzheimer representa un avance significativo en la detección y manejo de esta patología. Su implementación no solo beneficiará directamente a los pacientes y sus familias, sino que también tendrá un impacto positivo en el personal médico y en los sistemas de salud, al mejorar la precisión diagnóstica y reducir la subjetividad en la interpretación clínica. Además, el uso de imágenes PET con diferentes radiofármacos dirigidos a objetivos moleculares complementarios potencia la capacidad de detección temprana y la estratificación de los pacientes, permitiendo intervenciones más efectivas y personalizadas. Por lo tanto, en un contexto donde el envejecimiento poblacional es una realidad inminente, este tipo de innovaciones tecnológicas es fundamental para enfrentar los desafíos asociados con esta enfermedad neurodegenerativa.

Capítulo 4

Objetivos

4.1. Objetivo General

Desarrollar un sistema para el apoyo diagnóstico de la enfermedad de Alzheimer mediante algoritmos de inteligencia artificial (IA) entrenados con estudios de tomografía por emisión de positrones (PET).

4.2. Objetivos Específicos

1. Gestionar una base de datos de estudios PET provenientes de pacientes con diagnóstico de deterioro cognitivo, enfermedad de Alzheimer y sujetos de control.
2. Implementar algoritmos basados en redes neuronales convolucionales y máquinas de aprendizaje entrenados con estudios PET, que permitan identificar los distintos niveles de la enfermedad de Alzheimer.
3. Validar los algoritmos de aprendizaje mediante la evaluación con métricas estandarizadas para la clasificación.
4. Desarrollar e implementar una interfaz gráfica de visualización que permita a un usuario procesar estudios PET.

Capítulo 5

Marco de Referencia

5.1. Áreas Temáticas

Las áreas temáticas que abarca el presente proyecto son las siguientes:

- **Computers and Information Processing – Artificial Intelligence – Machine learning – Deep learning.**
- **Computers and Information Processing – Image Processing and Computer Vision – Biomedical image analysis – Medical image classification.**
- **Biomedical Engineering – Biomedical Signal and Image Processing – Medical imaging systems – Positron Emission Tomography (PET).**
- **Biomedical Engineering – Clinical Engineering – Diagnostic and therapeutic systems – Computer-aided diagnosis (CAD).**
- **Computers and Information Processing – Software – Human-computer interaction – Graphical user interfaces (GUI).**
- **Life Sciences – Neuroscience – Neuroimaging – Neurodegenerative diseases.**
- **Health Care – Public Health – Health informatics – Digital health systems.**

5.2. Marco Teórico

5.2.1. Enfermedad de Alzheimer y Deterioro Cognitivo

La enfermedad de Alzheimer (EA) es una patología neurodegenerativa progresiva, crónica e irreversible que constituye la causa más frecuente de demencia a nivel mundial, siendo responsable de aproximadamente entre el 60 % y el 80 % de los casos diagnosticados [36, 39]. Esta enfermedad se caracteriza por un deterioro gradual y sostenido de las funciones cognitivas, que afecta principalmente a la memoria episódica, el lenguaje, la capacidad de razonamiento, la orientación espacial y temporal, así como la habilidad para llevar a cabo actividades básicas e instrumentales de la vida diaria.

Desde el punto de vista clínico, la EA se manifiesta de manera progresiva, iniciándose generalmente con alteraciones leves de la memoria reciente, las cuales suelen pasar desapercibidas en

sus primeras etapas. A medida que la enfermedad avanza, se presentan déficits cognitivos más amplios, incluyendo dificultades en la comunicación, cambios conductuales, alteraciones en la toma de decisiones y, en fases avanzadas, una pérdida significativa de la autonomía funcional del paciente [40].

El deterioro cognitivo leve (Mild Cognitive Impairment, MCI) representa una condición intermedia entre el envejecimiento cognitivo normal y la demencia establecida. Los individuos con MCI presentan un declive cognitivo objetivable, superior al esperado para su edad y nivel educativo, pero que no interfiere de manera significativa con su independencia funcional [41]. No obstante, el MCI constituye un importante factor de riesgo para el desarrollo de la enfermedad de Alzheimer, ya que un porcentaje significativo de estos pacientes progresa hacia una demencia tipo Alzheimer en los años siguientes.

5.2.1.1. Fisiopatología

Desde una perspectiva neuropatológica, la enfermedad de Alzheimer se caracteriza principalmente por dos hallazgos patológicos fundamentales: la acumulación extracelular anómala de placas de beta-amiloide (Aβ) y la formación intracelular de ovillos neurofibrilares compuestos por proteína tau hiperfosforilada. Estas alteraciones desencadenan una cascada de eventos neurodegenerativos que incluyen disfunción sináptica, pérdida progresiva de neuronas y atrofia cerebral, afectando de manera predominante regiones como el hipocampo, la corteza temporal y la corteza parietal, las cuales desempeñan un papel esencial en los procesos de memoria, aprendizaje y funciones cognitivas superiores. En consecuencia, la progresión de estos cambios estructurales y funcionales se traduce en el deterioro cognitivo característico de la enfermedad, evidenciando la estrecha relación entre la carga neuropatológica y la manifestación clínica de la EA [42, 43].

5.2.1.2. Síntomas y manifestaciones clínicas

La enfermedad de Alzheimer se caracteriza por un curso clínico progresivo y heterogéneo, en el cual los síntomas cognitivos, conductuales y funcionales evolucionan de manera gradual a lo largo del tiempo. Esta progresión refleja el avance de los procesos neurodegenerativos subyacentes, que afectan inicialmente regiones cerebrales involucradas en la memoria y, posteriormente, se extienden a áreas responsables de funciones cognitivas superiores, comportamiento y control motor [36, 44].

Desde el punto de vista clínico, la EA suele dividirse en tres etapas principales, las cuales permiten describir la severidad del deterioro y orientar tanto el diagnóstico como el manejo terapéutico:

- **Etapa leve o prodrómica (Deterioro Cognitivo Leve, MCI):** Esta fase inicial se caracteriza por alteraciones sutiles pero persistentes de la memoria episódica, especialmente en la capacidad para retener información reciente. Los pacientes pueden presentar dificultades para encontrar palabras, fallos atencionales, problemas leves de planificación y una discreta desorientación temporal. Asimismo, son frecuentes los cambios emocionales, como la ansiedad, la depresión o la irritabilidad. A pesar de estas alteraciones, las actividades básicas de la vida diaria suelen conservarse, lo que diferencia esta etapa de una demencia establecida [41]. No

obstante, el MCI asociado al Alzheimer representa un estado de alto riesgo para la progresión hacia la demencia, lo que lo convierte en un objetivo clave para el diagnóstico temprano.

- **Etapa moderada:** En esta fase, el deterioro cognitivo se vuelve clínicamente evidente e interfiere de manera significativa con la autonomía del paciente. Se observa un empeoramiento notable de la memoria, desorientación en tiempo y espacio, dificultades en el lenguaje (afasia), alteraciones en el reconocimiento de objetos o personas (agnosia) y problemas en la ejecución de tareas complejas (apraxia). Desde el punto de vista conductual, pueden presentarse cambios marcados en la personalidad, apatía, agitación, conductas repetitivas y alteraciones del juicio. En esta etapa, los pacientes requieren supervisión frecuente para la realización de actividades instrumentales de la vida diaria [45].

- **Etapa severa:** La fase avanzada de la enfermedad se caracteriza por una pérdida profunda de las funciones cognitivas y una dependencia total del paciente. El lenguaje se ve gravemente afectado, pudiendo limitarse a frases cortas o desaparecer por completo. Se pierde la capacidad de reconocer a familiares cercanos y se presentan trastornos motores como rigidez, alteraciones de la marcha y dificultades para la deglución. Además, son frecuentes la incontinencia urinaria y fecal, así como complicaciones médicas asociadas con el deterioro físico generalizado. En esta etapa, el cuidado permanente es indispensable, y la enfermedad impacta de forma significativa tanto al paciente como a su entorno familiar y social [36].

Cabe resaltar que la progresión clínica de la enfermedad no siempre sigue un patrón uniforme entre los individuos, lo que introduce una alta variabilidad intersujeto. Esta heterogeneidad clínica representa uno de los principales retos diagnósticos, especialmente en las etapas tempranas, donde los síntomas pueden solaparse con el envejecimiento normal u otras patologías neurológicas. En este contexto, el uso de biomarcadores y técnicas de neuroimagen funcional, como la tomografía por emisión de positrones (PET), resulta fundamental para complementar la evaluación clínica y mejorar la detección temprana de la enfermedad [36].

5.2.2. Técnicas de diagnóstico de la Enfermedad de Alzheimer

El diagnóstico de la enfermedad de Alzheimer (EA) representa un desafío clínico complejo que requiere la integración de múltiples modalidades de evaluación. Tradicionalmente, el diagnóstico definitivo solo podía establecerse mediante confirmación histopatológica post-mortem, evidenciando la presencia de placas seniles de beta-amiloide y ovillos neurofibrilares de proteína tau hiperfosforilada [46]. Sin embargo, los avances científicos y tecnológicos de las últimas décadas han revolucionado la aproximación diagnóstica, permitiendo la detección temprana de biomarcadores patológicos y alteraciones estructurales, incluso en fases presintomáticas de la enfermedad [36].

En la actualidad, el diagnóstico clínico de la EA se fundamenta en criterios multidimensionales que incorporan evaluaciones neuropsicológicas, análisis de biomarcadores en líquido cefalorraquídeo (LCR) y técnicas avanzadas de neuroimagen [47]. Entre las modalidades de imagen más relevantes destacan la resonancia magnética (RM) y la tomografía por emisión de positrones (PET), las cuales

proporcionan información complementaria sobre los cambios estructurales, metabólicos y moleculares característicos de la patología. A continuación, se describen en detalle las principales técnicas de neuroimagen empleadas en el diagnóstico y seguimiento de la enfermedad de Alzheimer.

5.2.2.1. Resonancia Magnética (RM)

La resonancia magnética (RM) es una técnica de imagen médica no invasiva que permite obtener imágenes anatómicas detalladas del cuerpo humano, especialmente del sistema nervioso central. A diferencia de la tomografía computarizada (TC) o la tomografía por emisión de positrones (PET), la RM no utiliza radiación ionizante, sino campos magnéticos intensos y ondas de radiofrecuencia para generar imágenes de alta resolución de los tejidos blandos, lo que la convierte en una herramienta particularmente valiosa para el estudio del parénquima cerebral [48].

Principio físico

La RM se fundamenta en el fenómeno de la resonancia magnética nuclear, descubierto independientemente por Felix Bloch y Edward Purcell en 1946. En este proceso, el cuerpo del paciente es expuesto a un campo magnético intenso, típicamente entre 1.5 y 3 Tesla en aplicaciones clínicas, aunque existen equipos de hasta 7 Tesla para investigación, lo que provoca que los núcleos de hidrógeno (^{1}H), abundantes en las moléculas de agua y lípidos de los tejidos, se alineen con dicho campo magnético externo [49].

Posteriormente, se aplica un pulso de radiofrecuencia (RF) perpendicular al campo magnético principal, con una frecuencia específica denominada frecuencia de Larmor, que depende de la intensidad del campo magnético. Este pulso provoca que los protones absorban energía y cambien su orientación, desviándose del eje del campo magnético principal. Cuando cesa el pulso de RF, los protones regresan a su estado de equilibrio mediante dos procesos de relajación independientes: la relajación longitudinal (T1) y la relajación transversal (T2). Durante este retorno, los protones liberan energía en forma de señal electromagnética que es captada por las bobinas receptoras del equipo. Esta señal es procesada mediante transformadas de Fourier y algoritmos de reconstrucción de imagen, generando representaciones bidimensionales o tridimensionales de los tejidos [48, 50].

La intensidad de la señal captada varía según las propiedades fisicoquímicas de cada tejido (densidad de protones, tiempos de relajación T1 y T2), lo que permite diferenciar con alta precisión entre sustancia gris, sustancia blanca, líquido cefalorraquídeo y lesiones patológicas.

5.2.2.2. Tomografía por Emisión de Positrones (PET)

La tomografía por emisión de positrones (PET, del inglés Positron Emission Tomography) es una técnica de imagen molecular no invasiva que permite visualizar y cuantificar procesos fisiológicos, metabólicos y bioquímicos del organismo in vivo. A diferencia de las técnicas de imagen estructural como la RM o la TC, que proporcionan información anatómica, la PET ofrece información funcional y molecular con alta sensibilidad, aunque con menor resolución espacial. En el contexto del diagnóstico de enfermedades neurodegenerativas como el Alzheimer, la PET es especialmente valiosa porque permite detectar alteraciones bioquímicas y moleculares cerebrales en fases tempranas de

la enfermedad, incluso antes de que se manifiesten cambios estructurales evidentes en la resonancia magnética o síntomas clínicos significativos [51, 52].

Principio físico y fundamento de los estudios PET

El procedimiento PET se basa en la administración intravenosa de una sustancia radioactiva, denominada trazador o radiofármaco, que consiste en una molécula biológicamente activa marcada con un isótopo emisor de positrones. Esta molécula se distribuye en el organismo siguiendo sus vías metabólicas naturales y se acumula en los tejidos de interés según las propiedades fisicoquímicas y biológicas del compuesto. Los trazadores utilizados en neurología están marcados comúnmente con isótopos emisores de positrones de vida media corta, como el ^{18}F (Flúor-18, con vida media de 110 minutos), el ^{11}C (Carbono-11, con vida media de 20 minutos) o el ^{15}O (Oxígeno-15, con vida media de 2 minutos) [53].

Cuando el isótopo radiactivo decae, emite un positrón (partícula beta positiva, e^+) que recorre una distancia muy corta en el tejido (del orden de milímetros) hasta que colisiona con un electrón (e^-) del medio. Esta colisión produce un evento de aniquilación materia-antimateria que genera dos fotones gamma de 511 keV cada uno, emitidos simultáneamente en direcciones opuestas, formando un ángulo de aproximadamente 180°. El anillo de detectores de centelleo que rodea al paciente en el escáner PET está diseñado para captar coincidencias de estos pares de fotones dentro de una ventana temporal muy estrecha (típicamente 6-12 nanosegundos). La detección coincidente de ambos fotones permite localizar espacialmente el evento de aniquilación a lo largo de la línea que une ambos detectores, sin necesidad de colimación física. Mediante algoritmos avanzados de reconstrucción tomográfica (como la retroproyección filtrada o métodos iterativos), se obtiene una imagen tridimensional de la distribución del trazador en el cerebro, que refleja la actividad metabólica o la presencia de blancos moleculares específicos [51, 54].

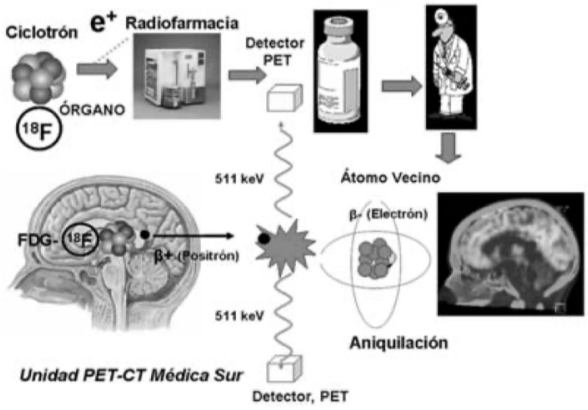


Figura 5.1: Principio físico de un estudio PET. El positrón emitido por el radioisótopo colisiona con un electrón, generando dos fotones gamma de 511 keV en direcciones opuestas que son detectados simultáneamente por el anillo de detectores. Adaptado de [1].

La principal ventaja de la PET radica en su capacidad para detectar cambios funcionales y

moleculares con alta sensibilidad, permitiendo identificar alteraciones patológicas en etapas presintomáticas. Sin embargo, presenta limitaciones en cuanto a la resolución espacial (típicamente 4-6 mm) y requiere la disponibilidad de un ciclotrón para la producción de radioisótopos de vida media corta, especialmente ^{11}C, lo que limita su uso a centros especializados [55].

Trazadores utilizados en la detección de la enfermedad de Alzheimer

En el diagnóstico de la EA, se utilizan tres categorías principales de trazadores, clasificados según la función fisiopatológica que evalúan. Cada uno proporciona información complementaria sobre diferentes aspectos de la cascada neurodegenerativa característica de la enfermedad [56]:

1. Trazadores metabólicos: ^{18}F-FDG (fluorodesoxiglucosa): La ^{18}F-FDG (2-fluoro-2-deoxi-D-glucosa) es el radiofármaco más ampliamente utilizado en la PET neurológica y representa el estándar de referencia para la evaluación del metabolismo cerebral de la glucosa. Estructuralmente, la ^{18}F-FDG es un análogo de la glucosa en el que un grupo hidroxilo en la posición 2 ha sido sustituido por un átomo de flúor-18. Esta molécula se transporta a través de la barrera hematoencefálica mediante los mismos transportadores de glucosa (GLUT) y es fosforilada intracelularmente por la hexoquinasa, formando ^{18}F-FDG-6-fosfato. Sin embargo, a diferencia de la glucosa-6-fosfato, la ^{18}F-FDG-6-fosfato no puede continuar la vía glucolítica ni ser desfosforilada eficientemente, quedando atrapada en el interior de las células metabólicamente activas. Por tanto, su acumulación tisular refleja directamente la tasa de consumo de glucosa y, por extensión, la actividad metabólica celular [57, 58].

En condiciones fisiológicas, el cerebro exhibe un patrón de captación de ^{18}F-FDG relativamente homogéneo en la corteza cerebral, con mayor actividad en la corteza cingulada, los ganglios basales y el cerebelo. En contraste, los pacientes con EA presentan un patrón característico de hipometabolismo que afecta preferentemente a regiones específicas y que evoluciona de manera predecible a lo largo de la progresión de la enfermedad [52, 59]:

- En etapas tempranas (deterioro cognitivo leve), se observa hipometabolismo en el cíngulo posterior y el precúneo, regiones que forman parte de la red neuronal por defecto.
- En etapas moderadas, el hipometabolismo se extiende a la corteza parietotemporal bilateral, afectando áreas asociativas multimodales cruciales para la memoria y el procesamiento visuoespacial.
- En fases avanzadas, las regiones frontales, incluida la corteza prefrontal dorsolateral y la corteza temporal medial (incluido el hipocampo), también muestran una reducción metabólica significativa.

2. Trazadores para beta-amiloide: ^{11}C-PIB, ^{18}F-Florbetapir, ^{18}F-Flutemetamol, ^{18}F-Florbetaben: El desarrollo de trazadores específicos para la visualización in vivo de placas de beta-amiloide (Aβ) ha representado uno de los avances más significativos en la neuroimagen de la EA. Estos radiofármacos se unen con alta afinidad y especificidad a los agregados fibrilares de beta-amiloide depositados en el espacio extracelular del parénquima cerebral, permitiendo cuantificar la

carga amiloide total. La utilidad clínica de estos trazadores radica en su capacidad para identificar sujetos con acumulación patológica de amiloide en fases presintomáticas (amiloidosis cerebral precoz), incluso décadas antes del inicio de los síntomas cognitivos, lo que resulta fundamental para el diseño de estrategias terapéuticas preventivas [60, 56].

La captación de estos trazadores es particularmente elevada en áreas corticales de asociación, como la corteza frontal, el precúneo, el cíngulo posterior y la corteza temporal, mientras que es mínima en el cerebelo, que habitualmente se utiliza como región de referencia por su baja carga de amiloide, incluso en pacientes con EA [60].

Es importante destacar que la positividad para amiloide en PET no es sinónimo de enfermedad de Alzheimer clínica, ya que aproximadamente el 20-30 % de los adultos mayores cognitivamente normales presentan acumulación significativa de amiloide cerebral (amiloidosis presintomática), lo que se considera un factor de riesgo para desarrollar deterioro cognitivo en el futuro [36, 61]. Por otro lado, la ausencia de señal amiloide en PET tiene un alto valor predictivo negativo, permitiendo excluir la EA como causa del deterioro cognitivo con alta certeza [62].

3. Trazadores para proteína tau: ^{18}F-Flortaucipir (AV-1451), ^{18}F-MK-6240, ^{18}F-PI-2620: La proteína tau hiperfosforilada, que se agrega formando ovillos neurofibrilares intracelulares, constituye el otro componente neuropatológico fundamental de la EA. Recientemente, se han desarrollado trazadores PET de segunda generación capaces de unirse específicamente a los agregados de tau patológica, permitiendo su visualización in vivo. A diferencia de la deposición de amiloide, que sigue un patrón relativamente difuso y no se correlaciona estrechamente con la severidad clínica, la acumulación de tau muestra una distribución topográfica que sigue la progresión jerárquica descrita por Braak y Braak, iniciándose en estructuras del lóbulo temporal medial (corteza entorrinal transentorrinal) y extendiéndose progresivamente hacia regiones neocorticales temporales, parietales y frontales [46].

5.2.2.3. Evaluaciones cognitivas y neuropsicológicas

Además de las técnicas de neuroimagen y los biomarcadores biológicos, el diagnóstico de la enfermedad de Alzheimer se apoya de manera fundamental en evaluaciones cognitivas y neuropsicológicas estandarizadas. Estas pruebas permiten cuantificar de forma objetiva el rendimiento cognitivo del paciente y detectar alteraciones funcionales que reflejan el impacto clínico de los cambios neurodegenerativos subyacentes.

Las evaluaciones cognitivas constituyen, en muchos casos, el primer punto de contacto diagnóstico, ya que permiten identificar déficits en dominios como la memoria, la atención, el lenguaje y las funciones ejecutivas, incluso antes de que las alteraciones estructurales o moleculares sean claramente visibles en estudios de imagen. En la práctica clínica moderna, estas pruebas se utilizan de manera complementaria a la RM y la PET, proporcionando un marco funcional que facilita la interpretación de los hallazgos neurobiológicos y la estratificación del estadio de la enfermedad [36].

Entre las herramientas más utilizadas se encuentra el *Mini-Mental State Examination* (MMSE), debido a su simplicidad, reproducibilidad y amplia validación en poblaciones con deterioro cognitivo y enfermedad de Alzheimer.

5.2.2.4. Mini-Mental State Examination (MMSE)

El *Mini-Mental State Examination* (MMSE) es una prueba neuropsicológica breve y estructurada, ampliamente utilizada para la evaluación del estado cognitivo global en entornos clínicos y de investigación. Fue desarrollada originalmente por Folstein y colaboradores como una herramienta de cribado para detectar deterioro cognitivo y demencia, y desde entonces se ha convertido en uno de los instrumentos más empleados en el estudio de la enfermedad de Alzheimer [63].

El MMSE consta de un puntaje máximo de 30 puntos y evalúa múltiples dominios cognitivos, incluyendo la orientación temporal y espacial, la memoria inmediata y diferida, la atención y el cálculo, el lenguaje (comprensión, denominación y repetición), y las habilidades visuoconstructivas. La administración de la prueba es rápida, generalmente inferior a 10 minutos, lo que facilita su aplicación en la práctica clínica rutinaria.

Desde el punto de vista de la interpretación clínica, los puntajes del MMSE se asocian de forma directa con el grado de deterioro cognitivo. De manera general, puntuaciones entre 27 y 30 se consideran indicativas de una función cognitiva preservada; valores entre 21 y 26 suelen corresponder a deterioro cognitivo leve; puntuaciones entre 11 y 20 reflejan deterioro cognitivo moderado; y valores entre 0 y 10 se asocian con deterioro cognitivo severo. No obstante, la interpretación de estos rangos debe realizarse considerando factores como la edad y el nivel educativo del paciente [64].

5.2.3. Alzheimer's Disease Neuroimaging Initiative (ADNI)

La *Alzheimer's Disease Neuroimaging Initiative* (ADNI) es una de las iniciativas de investigación más reconocidas y ampliamente utilizadas en el estudio de la enfermedad de Alzheimer. Lanzada en 2004 como un estudio observacional longitudinal multicéntrico, ADNI fue concebida con el objetivo primario de validar biomarcadores que permitan la detección temprana, el seguimiento de la progresión y la evaluación de nuevas intervenciones terapéuticas en la EA [27].

5.2.3.1. Estructura y gestión de datos

ADNI es administrada por el Laboratorio de Neuroimagen (Laboratory of Neuro Imaging, LONI) del Instituto de Neuroimagen e Informática Mark y Mary Stevens de la Universidad del Sur de California (USC). La iniciativa ha establecido una infraestructura robusta para la recopilación, procesamiento, control de calidad, almacenamiento y distribución de datos clínicos y de neuroimagen de manera estandarizada y accesible a la comunidad científica internacional.

Una característica distintiva de ADNI es su política de datos abiertos, que permite a investigadores calificados de todo el mundo acceder libremente a los datos, previa solicitud y aprobación del comité directivo. Esta filosofía ha democratizado el acceso a información de alta calidad y ha catalizado la producción científica, generando más de 3,000 publicaciones derivadas de análisis de datos ADNI [65].

5.2.3.2. Población y diseño del estudio

ADNI ha reclutado participantes de aproximadamente 60 centros clínicos distribuidos en Estados Unidos y Canadá. Los participantes se clasifican en tres grupos clínicos principales: adultos mayores cognitivamente normales (CN), pacientes con deterioro cognitivo leve (MCI) y pacientes con enfermedad de Alzheimer en etapa leve (AD). Todos los participantes son sometidos a evaluaciones clínicas, neuropsicológicas y de neuroimagen estandarizadas en intervalos regulares que van desde 6 hasta 24 meses, dependiendo de la fase del estudio y el grupo clínico [66].

5.2.3.3. Modalidades de datos

ADNI ha establecido protocolos estandarizados para la adquisición de datos multimodales, garantizando la comparabilidad entre centros.

- **Resonancia Magnética (MRI):** Imágenes estructurales ponderadas en T1 de alta resolución utilizando secuencias MPRAGE, optimizadas para análisis volumétrico. En fases avanzadas, se incorporaron secuencias DTI y resonancia magnética funcional en estado de reposo (rs-fMRI).
- **Tomografía por Emisión de Positrones (PET):** Protocolos estandarizados para múltiples trazadores, incluyendo ^{18}F-FDG para metabolismo cerebral, trazadores de amiloide (^{18}F-florbetapir, ^{18}F-florbetaben, ^{18}F-flutemetamol) y trazadores de tau (^{18}F-flortaucipir).
- **Biomarcadores de fluidos:** Obtención de líquido cefalorraquídeo (LCR) y muestras sanguíneas para la cuantificación de beta-amiloide 1-42, proteína tau total, tau fosforilada y genotipificación de APOE.
- **Evaluaciones neuropsicológicas:** Batería comprensiva que evalúa memoria episódica, función ejecutiva, lenguaje, atención y velocidad de procesamiento, junto con escalas de funcionalidad global (CDR, ADAS-Cog).

5.2.4. Inteligencia Artificial (IA)

La inteligencia artificial (IA) se refiere, de forma general, a cualquier sistema computacional capaz de realizar tareas que normalmente requieren inteligencia humana, tales como la percepción visual, el reconocimiento de voz, la toma de decisiones y la traducción de idiomas [67]. En el contexto médico y científico actual, la IA ha emergido como una herramienta transformadora con aplicaciones que abarcan desde el diagnóstico asistido por computadora hasta la predicción de la evolución clínica y la personalización de tratamientos [68].

Es fundamental distinguir entre los diferentes niveles y enfoques dentro del campo de la IA, particularmente entre los conceptos de aprendizaje automático (machine learning) y aprendizaje profundo (deep learning), ya que estos términos, aunque relacionados, representan paradigmas distintos con características y capacidades específicas.

5.2.4.1. Aprendizaje Automático (Machine Learning)

El aprendizaje automático (ML, del inglés Machine Learning) constituye un subcampo de la inteligencia artificial que se enfoca en el desarrollo de algoritmos y modelos estadísticos que permiten a los sistemas computacionales aprender y mejorar su desempeño en una tarea específica a partir de la experiencia, sin ser explícitamente programados para cada escenario posible [69]. A diferencia de los sistemas tradicionales basados en reglas predefinidas, tales como estructuras condicionales del tipo *if-then-else*, los algoritmos de ML infieren patrones y relaciones directamente de los datos.

Fundamentos del machine learning

En este mismo orden, el proceso de aprendizaje automático se basa en el uso de algoritmos que reciben como entrada un conjunto de datos con características y etiquetas, lo que genera una función que mapea la entrada a la salida. Siendo esta función la que puede ser utilizada para predecir el resultado de nuevas observaciones.

En términos generales, un modelo de machine learning se entrena ajustando sus parámetros para minimizar una función de pérdida que cuantifica el error entre las predicciones del modelo y los valores reales. En otras palabras, el aprendizaje puede ser:

- **Supervisado:** Es cuando se dispone de etiquetas, lo que incluye tareas como clasificación y regresión.
- **No supervisado:** Es útil cuando no hay etiquetas y se busca descubrir estructuras ocultas.
- **Semi-supervisado o por refuerzo:** Se utiliza para combinaciones o aprendizaje mediante interacción.

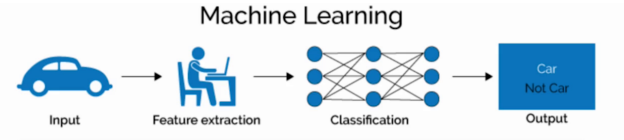


Figura 5.2: Diagrama Explicativo de Machine Learning [2].

5.2.4.2. Aprendizaje Profundo (Deep Learning)

El aprendizaje profundo (DL, del inglés Deep Learning) representa el estado del arte dentro del aprendizaje automático y se basa fundamentalmente en el uso de redes neuronales artificiales profundas, es decir, arquitecturas con múltiples capas de procesamiento jerárquico que permiten al modelo aprender representaciones de datos de complejidad creciente [70]. La característica distintiva del DL es su capacidad para realizar aprendizaje de representaciones (representation learning), donde el modelo no solo aprende a mapear entradas a salidas, sino que también descubre automáticamente las características relevantes de los datos de entrada sin necesidad de ingeniería manual de características.

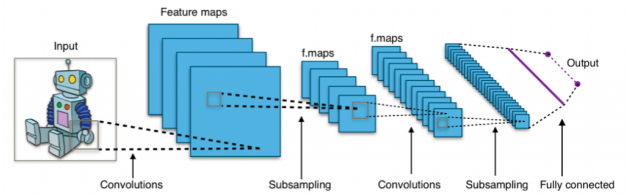


Figura 5.3: Diagrama Explicativo de Deep Learning [3].

5.2.4.3. Redes Neuronales Convolucionales (CNN)

Las redes neuronales convolucionales (CNN) son un tipo de red neuronal profunda diseñada específicamente para el procesamiento de datos con estructura de cuadrícula, tales como imágenes. Estas fueron popularizadas por su éxito en tareas de visión por computadora, como clasificación de imágenes, detección de objetos y segmentación semántica [70].

Funcionamiento general de una CNN Principalmente, las CNN utilizan capas especializadas que extraen automáticamente características espaciales de los datos de entrada. Por lo tanto, su funcionamiento se basa en tres componentes fundamentales:

1. **Capas convolucionales**
2. **Capas de activación**
3. **Capas de agrupamiento o pooling**
4. **Capas totalmente conectadas**

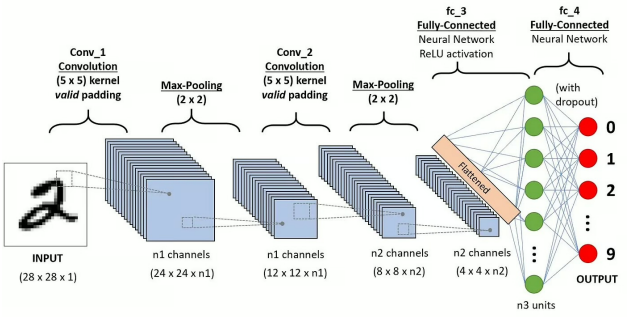


Figura 5.4: Arquitectura de una CNN [4].

Entrenamiento de una CNN

De este modo, las CNN se entrenan utilizando aprendizaje supervisado. En el cual, a partir de un conjunto de datos etiquetado, se ajustan los pesos de los filtros mediante el algoritmo de retropropagación y la optimización por descenso de gradiente. Siendo la entropía cruzada la función de pérdida comúnmente usada en clasificación multiclase:

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

donde y_i es la etiqueta verdadera, que está codificada en one-hot y $\hat{y}_i$ es la probabilidad predicha para la clase i.

5.2.5. Transferencia de aprendizaje (Transfer Learning)

La *transferencia de aprendizaje* (*transfer learning*) es una técnica ampliamente utilizada en el campo del aprendizaje automático y del aprendizaje profundo, cuyo objetivo es aprovechar el conocimiento previamente adquirido por un modelo entrenado en una tarea o dominio específico y aplicarlo a la resolución de un nuevo problema relacionado. En lugar de entrenar un modelo desde cero, este enfoque permite reutilizar representaciones, patrones y parámetros aprendidos con anterioridad, lo cual resulta especialmente útil cuando se dispone de conjuntos de datos limitados, situación frecuente en aplicaciones reales y, particularmente, en el ámbito médico [71, 72].

En el contexto de las redes neuronales convolucionales (CNN), la transferencia de aprendizaje se fundamenta en el hecho de que las capas iniciales de estas redes aprenden características generales de las imágenes, como bordes, esquinas, texturas y patrones básicos, que suelen ser comunes entre distintas tareas de visión por computador. Por esta razón, dichas capas pueden reutilizarse eficazmente en nuevos problemas sin necesidad de ser entrenadas nuevamente. En contraste, las capas más profundas capturan características de alto nivel, más específicas de la tarea original, por lo que suelen requerir adaptación al nuevo dominio mediante un proceso conocido como *fine-tuning* [73].

Esta estrategia ha demostrado ser particularmente efectiva en el análisis de imágenes médicas, donde la disponibilidad de grandes volúmenes de datos etiquetados es limitada debido a restricciones técnicas, éticas y económicas. por lo tanto, mediante el uso de transferencia de aprendizaje, es posible mejorar la capacidad de generalización de los modelos, reducir el riesgo de sobreajuste y disminuir de manera significativa el tiempo y los recursos computacionales necesarios para el entrenamiento [74].

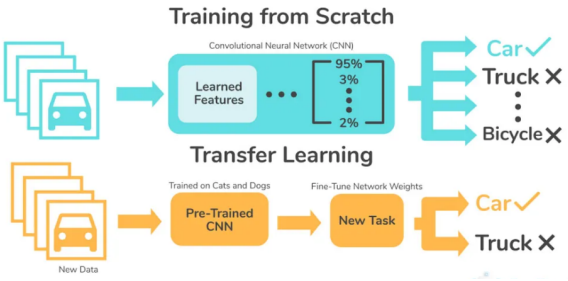


Figura 5.5: Diagrama Explicativo de Transfer Learning [5].

5.2.5.1. Modelos híbridos CNN + ML

Los modelos híbridos en inteligencia artificial son arquitecturas que combinan diferentes enfoques o algoritmos con el fin de aprovechar las fortalezas particulares de cada técnica y compensar sus debilidades. En el ámbito biomédico, los modelos híbridos han ganado creciente popularidad debido a su capacidad para integrar información multimodal, como imágenes, datos clínicos y características demográficas, proporcionando una aproximación más robusta, personalizada y explicativa en tareas de clasificación o predicción diagnóstica [75].

Arquitectura general de un modelo híbrido

La estructura de un modelo híbrido puede variar dependiendo del tipo de datos, pero generalmente presenta tres componentes clave:

1. **Módulo de extracción de características**: Se encarga de transformar los datos de entrada en un conjunto de características numéricas relevantes. En este contexto, este módulo suele estar conformado por una CNN, que extrae patrones espaciales y texturales complejos a partir de imágenes volumétricas o planos bidimensionales [76].

2. **Módulo de integración de datos tabulares**: En paralelo, los datos no imagenológicos como variables clínicas, son procesados mediante técnicas clásicas de preprocesamiento, como la normalización o imputación, para generar vectores de entrada compatibles.

3. **Módulo de clasificación final**: Las salidas del extractor CNN, como vectores de activación, pueden conectarse directamente a un clasificador de machine learning, como un modelo de *K-Nearest Neighbors* (KNN), máquinas de vectores de soporte (SVM) o árboles de decisión (RF), en lugar de utilizar capas densas tradicionales. Alternativamente, ambas fuentes, es decir, imagen + datos clínicos, pueden concatenarse para alimentar un clasificador mixto [77].

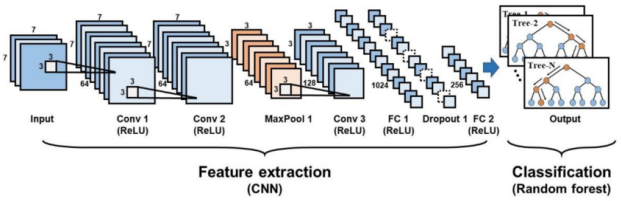


Figura 5.6: Diagrama Explicativo de los Modelos Híbridos [6].

5.2.6. Métricas de evaluación de modelos de clasificación

Las métricas de evaluación son herramientas fundamentales en el aprendizaje automático supervisado, ya que permiten cuantificar de manera objetiva el rendimiento de un modelo una vez finalizado su proceso de entrenamiento. En el contexto del diagnóstico asistido por inteligencia artificial, el uso de estas métricas resulta esencial para determinar la utilidad clínica de un sistema,

mediante la comparación entre las predicciones realizadas por el modelo y las etiquetas reales de los datos [78].

En tareas de clasificación multiclase, como la detección de la enfermedad de Alzheimer en distintos estadios, no es suficiente evaluar únicamente la proporción total de aciertos. En su lugar, se emplean métricas que permiten analizar el comportamiento del modelo para cada clase individual, lo cual es especialmente relevante en escenarios clínicos donde los errores pueden tener consecuencias significativas. El Cuadro5.1muestra las métricas de evaluación implementadas dentro de este proyecto.

Cuadro 5.1: Métricas de evaluación utilizadas en los modelos de clasificación

Métrica	**Qué mide**	**Interpretación**
Accuracy	Proporción total de predicciones correctas	Indica el rendimiento global del modelo, pero puede ser engañosa en conjuntos de datos desbalanceados.
Precisión	Exactitud de las predicciones positivas	Mide qué tan confiables son las predicciones positivas realizadas por el modelo.
Sensibilidad (Recall)	Capacidad para detectar casos positivos reales	Evalúa qué tan bien el modelo identifica a los pacientes enfermos; métrica crítica en aplicaciones clínicas.
Especificidad	Capacidad para identificar correctamente los casos negativos	Indica qué tan bien el modelo evita clasificar erróneamente a sujetos sanos.
F1-score	Balance entre precisión y sensibilidad	Resume el desempeño del modelo cuando existe desbalance entre clases.
AUC-ROC	Capacidad de discriminación entre clases	Mide qué tan bien el modelo separa las clases independientemente del umbral de decisión.
Matriz de confusión	Distribución de aciertos y errores por clase	Permite analizar en qué clases el modelo comete más errores.
Promedio macro	Promedio equitativo entre clases	Asigna el mismo peso a todas las clases, útil cuando interesa el desempeño balanceado.
Promedio micro	Promedio ponderado por número de muestras	Refleja el desempeño global considerando el tamaño de cada clase.

5.2.7. Plataforma de Visualización y Diseño de Interfaz Gráfica

La visualización eficaz constituye un componente esencial en los sistemas médicos asistidos por inteligencia artificial (IA), especialmente en el diagnóstico de enfermedades neurodegenerativas como el Alzheimer. Por su parte, estas plataformas deben garantizar la representación precisa de los

resultados generados por los modelos de IA, facilitando su interpretación por parte de profesionales clínicos.

Para ello, es fundamental implementar interfaces gráficas centradas en el usuario (GUI), diseñadas bajo principios de usabilidad, accesibilidad y experiencia de usuario (UX). Según [79], el diseño centrado en el usuario mejora la seguridad, la eficiencia y la aceptación de los sistemas clínicos. Además, las guías de la [80] establecen directrices clave para el diseño interactivo de sistemas enfocados en las necesidades del usuario final. En este contexto, herramientas como **Kivy** (Python), **PyQt** o entornos web basados en **Dash** y **Streamlit** permiten integrar visualizaciones interactivas de datos, imágenes médicas y controles funcionales, siendo ampliamente utilizadas en prototipos de sistemas de apoyo clínico [81].

5.3. Trabajos Relacionados

El diagnóstico temprano de la enfermedad de Alzheimer (EA) ha sido objeto de diversos estudios en los últimos años, particularmente en el área de la inteligencia artificial, el aprendizaje automático y el procesamiento de imágenes médicas. A continuación, se describen seis trabajos clave que abordan esta problemática desde distintos enfoques metodológicos y técnicos, sirviendo de referencia para el desarrollo del presente proyecto.

En primer lugar, el trabajo de Solanas Sanz [18] propone un sistema de aprendizaje profundo basado en CNNs para predecir el diagnóstico de la enfermedad de Alzheimer a partir de imágenes 18F-FDG PET preprocesadas del repositorio ADNI, con el objetivo principal de clasificar a los pacientes en tres grupos: cognitivamente normales (CN), con deterioro cognitivo leve (MCI) y con diagnóstico de Alzheimer (AD). Cabe resaltar que el autor busca emplear arquitecturas como InceptionV3 e InceptionResNetV2, logrando una precisión del 84,6 % y un AUC de 0,89, anticipando el diagnóstico clínico hasta 66 meses. Aquí, su enfoque se destaca por incorporar técnicas de transferencia de aprendizaje, lo cual mejora el rendimiento ante conjuntos de datos reducidos, lo que valida la eficacia del aprendizaje profundo en la identificación de patrones tempranos de deterioro cognitivo. Finalmente, este estudio resalta también la importancia de emplear imágenes PET como biomarcadores en las fases iniciales del Alzheimer, dado que permiten detectar alteraciones en el metabolismo de la glucosa cerebral incluso antes de la aparición de síntomas clínicos. Asimismo, subraya la viabilidad técnica de utilizar CNNs para detectar patrones visuales complejos en las imágenes médicas, reduciendo así la dependencia del análisis experto tradicional.

Por su parte, Ding et al. [16] desarrollaron un modelo de aprendizaje profundo específicamente diseñado para predecir el diagnóstico de EA utilizando FDG-PET cerebral. Su red neuronal convolucional 3D alcanzó una exactitud del 72.3 % en la clasificación binaria EA vs. controles normales (CN). Los autores demostraron que su modelo puede identificar patrones de hipometabolismo regional característicos de la EA, con especial énfasis en regiones temporoparietales y del cíngulo posterior. Este trabajo es particularmente relevante ya que utiliza únicamente imágenes PET, sin requerir datos multimodales adicionales.

Ahora bien, Fernández Cobas [19] adoptó un enfoque estadístico y de machine learning para analizar factores de riesgo asociados a la EA, tales como la edad, los hábitos de vida y las varia-

bles genéticas. En donde se implementaron modelos como K-Nearest Neighbors (KNN) y regresión logística, que, si bien mostraron precisiones individuales menores al 60 %, mejoraron notablemente cuando se integraron con pruebas clínicas. Reafirmando así el potencial diagnóstico de variables no imagenológicas en etapas tempranas de la enfermedad.

Por otro lado, tenemos a Aguilar Obregón [20], quien diseñó un algoritmo de aprendizaje autónomo basado en variables sociodemográficas y clínicas como edad, educación, presión arterial, consumo de sustancias y genotipo APOE. Cabe resaltar que el estudio implementa tres algoritmos principales, desde regresión logística, K-Nearest Neighbors (KNN) hasta Naive Bayes. Cada modelo fue entrenado y validado sobre un conjunto de datos estructurado y normalizado. Así pues, en las pruebas realizadas, el modelo basado en KNN alcanzó una precisión del 83.3 % en la predicción de casos de Alzheimer, mientras que la regresión logística mostró mayor estabilidad en términos de interpretabilidad de resultados. Por su parte, Naive Bayes presentó resultados aceptables, aunque con menor precisión en comparación. Finalmente, el algoritmo es validado con múltiples casos y se plantea su posible integración como servicio web, resaltando la utilidad práctica del sistema. Este enfoque apunta a un modelo accesible y escalable, ideal para contextos donde las imágenes médicas son limitadas.

Asimismo, en un estudio reciente, Rogeau et al [82] desarrollaron y validaron un modelo de red neuronal convolucional tridimensional (3D CNN) inspirado en la arquitectura VGG16, con el objetivo de clasificar imágenes cerebrales obtenidas por tomografía por emisión de positrones (PET) utilizando el radiofármaco 18F-FDG. El preprocesamiento de las imágenes incluyó normalización espacial al template ICBM152, segmentación para excluir regiones no cerebrales (hueso, aire, tejido blando) y normalización de intensidades voxel-wise entre 0 y 1. Se aplicaron técnicas de aumento de datos en tiempo real, como rotaciones aleatorias, flips y translaciones tridimensionales, para reducir el sobreajuste. La selección de hiperparámetros se realizó mediante optimización bayesiana, y el entrenamiento se efectuó usando validación cruzada de 5 pliegues. El rendimiento del modelo se evaluó en el conjunto de prueba y se comparó directamente con la interpretación visual realizada por tres médicos especialistas en medicina nuclear, sin acceso a datos clínicos. El modelo alcanzó una precisión del 89.8 % en la clasificación de las imágenes, superando al consenso de los especialistas, cuya precisión fue del 69.5 %. Los valores del área bajo la curva ROC (AUC) fueron 93.3 % para AD, 95.3 % para FTD y 99.9 % para CN, siendo la clasificación de sujetos normales la más precisa, con una sensibilidad del 100 % y especificidad del 97 %.

Finalmente, Liu et al. [83] implementaron una arquitectura ResNet3D para clasificación multiclase (EA/MCI/CN) utilizando imágenes de FDG-PET. Su modelo en cascada de redes neuronales convolucionales alcanzó una exactitud del 68.9 % en la tarea de clasificación de tres clases. Los autores identificaron que la clase MCI presenta una mayor dificultad de clasificación debido a su heterogeneidad clínica, ya que algunos pacientes progresan a EA mientras que otros permanecen estables. Este hallazgo es consistente con la naturaleza transitoria del MCI como estado intermedio entre el envejecimiento normal y la demencia.

Capítulo 6

Materiales y Métodos

6.1. Base de datos

6.1.1. Descripción del dataset de los estudios PET

Para el desarrollo de este trabajo se utilizaron imágenes de tomografía por emisión de positrones (PET) provenientes de la base de datos del *Alzheimer's Disease Neuroimaging Initiative* (ADNI), la cual incluye diversos tipos de radiofármacos utilizados en estudios PET, cada uno diseñado para visualizar diferentes procesos patológicos característicos de la enfermedad de Alzheimer. Aunque ADNI no clasifica formalmente las imágenes PET según el tipo de radiofármaco en grupos diagnósticos predefinidos, para los fines de este estudio y con el objetivo de facilitar el análisis automatizado mediante técnicas de aprendizaje profundo, se ha optado por organizar los radiofármacos en tres grandes categorías funcionales. Esta agrupación se basa en el mecanismo biológico que cada trazador permite visualizar y en las regiones cerebrales que tienden a resaltar, lo cual permite estructurar mejor los datos y diseñar modelos de aprendizaje automático más específicos para cada modalidad de imagen.

6.1.1.1. Clasificación de radiofármacos por grupo funcional

La clasificación propuesta organiza los trazadores en tres grupos principales: Amiloide, Tau y Metabólico. A continuación, se describe cada uno de ellos:

1. Grupo Amiloide

Este grupo está conformado por radiofármacos diseñados específicamente para detectar placas de proteína β-amiloide en el tejido cerebral. Las placas amiloides constituyen uno de los primeros signos patológicos de la enfermedad de Alzheimer, tendiendo a acumularse en etapas tempranas de la enfermedad, incluso décadas antes de la aparición de síntomas clínicos [84, 85]. Estos depósitos proteicos se distribuyen predominantemente en regiones corticales como la corteza frontal, el precuneus y áreas parietales y temporales.

2. Grupo Tau

Este grupo agrupa los radiofármacos que permiten detectar ovillos neurofibrilares formados por proteínas Tau hiperfosforiladas. A diferencia de las placas amiloides, los ovillos de Tau aparecen en etapas más avanzadas de la enfermedad y muestran una mayor correlación espacial y temporal con los síntomas clínicos del deterioro cognitivo [86, 87]. La distribución de estos depósitos sigue típicamente un patrón de progresión que inicia en regiones mediales del lóbulo temporal y se extiende gradualmente hacia áreas neocorticales.

3. Grupo Metabólico

Este grupo está representado principalmente por el radiofármaco [^{18}F]FDG (fluorodeoxiglucosa), un análogo de la glucosa que permite evaluar el metabolismo cerebral regional. A diferencia de los grupos anteriores, FDG no detecta depósitos proteicos específicos, sino que ofrece una medida indirecta de la actividad neuronal y el consumo energético celular [59].

El patrón de hipometabolismo observado mediante FDG-PET en la enfermedad de Alzheimer típicamente incluye una reducción de la captación en regiones temporoparietales, la corteza cingulada posterior y el precuneus, mientras que las áreas sensoriomotoras, visuales primarias y ganglios basales tienden a preservarse hasta etapas avanzadas. Este trazador es particularmente útil para la diferenciación entre distintos tipos de demencia, como la enfermedad de Alzheimer y la demencia frontotemporal, dado que cada una presenta patrones metabólicos característicos [88].

6.1.1.2. Organización y distribución del dataset

El conjunto de datos final utilizado en este estudio comprende un total de 5,673 imágenes PET, distribuidas según el tipo de radiofármaco empleado. El Cuadro 6.1 presenta la distribución detallada de los estudios por grupo y clase específica de radiofármacos.

Cuadro 6.1: Distribución de estudios por grupo funcional y clase de radiofármaco

Grupo	Clase de Radiofármaco	N° de estudios
Amiloide	GRUPO_AMILOIDE-AD	117
	GRUPO_AMILOIDE-CN	168
	GRUPO_AMILOIDE-MCI	498
Metabólico	GRUPO_METABOLICO-AD	155
	GRUPO_METABOLICO-CN	235
	GRUPO_METABOLICO-MCI	334
Tau	GRUPO_TAU-AD	285
	GRUPO_TAU-CN	2,479
	GRUPO_TAU-MCI	1,402
Total		**5,673**

donde AD corresponde a pacientes diagnosticados con enfermedad de Alzheimer, CN a individuos cognitivamente normales (controles) y MCI a pacientes con deterioro cognitivo leve.

La Figura 6.1 muestra la proporción de estudios correspondiente a cada grupo funcional. Se observa que el Grupo Tau representa la mayor parte del dataset, con un 73.44 % del total de imágenes (4,166 estudios), seguido por el Grupo Amiloide con un 13.80 % (783 estudios) y el Grupo Metabólico con un 12.76 % (724 estudios).

La Figura 6.2 presenta el conteo absoluto de estudios para cada combinación de grupo funcional y categoría diagnóstica. Se aprecia una distribución heterogénea entre las diferentes clases, siendo la categoría GRUPO_TAU-CN la más numerosa, con 2,479 estudios, mientras que la categoría GRUPO_METABOLICO-AD presenta el menor número, con 155 estudios. Esta disparidad en la

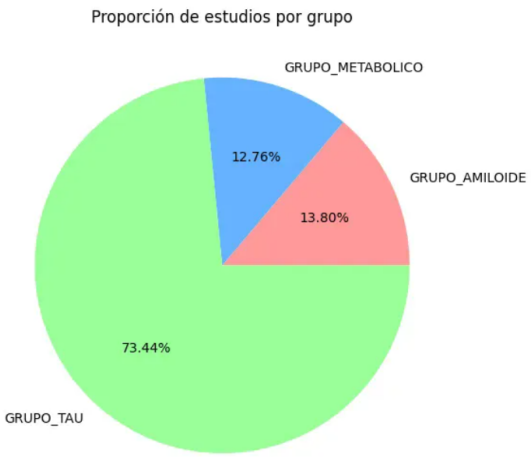


Figura 6.1: Proporción de estudios por grupo funcional de radiofármacos

cantidad de muestras por clase representa un desafío importante para el entrenamiento de modelos de aprendizaje automático, requiriendo la implementación de técnicas de balanceo de clases o estrategias de aumento de datos.

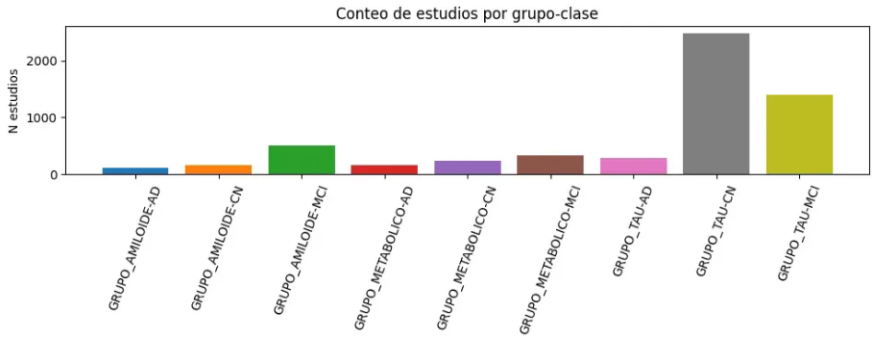


Figura 6.2: Conteo de estudios por grupo funcional y categoría diagnóstica

6.1.2. Base de datos de las Variables sociodemográficas y clínicas

6.1.2.1. Características generales del dataset sociodemográfico

El conjunto de datos sociodemográficos y clínicos utilizados en este estudio comprende un total de 4,617 sujetos con información completa sobre variables demográficas, clínicas y cognitivas. La distribución por género muestra una ligera predominancia femenina, con 2,342 mujeres (50.73 %) y 2,275 hombres (49.27 %), lo cual es consistente con la mayor prevalencia de la enfermedad de Alzheimer en la población femenina reportada en la literatura epidemiológica.

La edad promedio de la población total es de 71.85 años, con un rango que oscila entre 54

y 91 años. Al analizar la distribución por género, se observa que los hombres presentan una edad promedio ligeramente superior, con 73.15 años, en comparación con las mujeres, que tienen una edad promedio de 70.58 años, con rangos de edad de 54 a 91 años y de 50 a 91 años, respectivamente.

En cuanto a la distribución por grupo diagnóstico, el Cuadro 6.2 presenta el desglose detallado de la muestra. El grupo de Deterioro Cognitivo Leve (MCI) constituye la categoría más numerosa, con 2,181 sujetos (47.24 % del total), seguido por el grupo de Cognitivamente Normal (CN) con 1,850 sujetos (40.07 %), y finalmente, el grupo de Enfermedad de Alzheimer (AD) con 586 sujetos (12.69 %). Esta distribución refleja el enfoque de ADNI en el estudio de las etapas tempranas y de transición de la enfermedad, particularmente el deterioro cognitivo leve, que representa un período crítico para intervenciones terapéuticas.

Cuadro 6.2: Distribución de sujetos por grupo diagnóstico

Grupo	**Cantidad**	**Porcentaje**	**Edad Promedio (DE)**
AD	586	12.69 %	74.74 años (8.09)
CN	1,850	40.07 %	70.51 años (7.10)
MCI	2,181	47.24 %	72.21 años (7.60)
Total	**4,617**	**100.00 %**	**71.85 años (7.52)**

El análisis de edad por grupo diagnóstico revela un patrón esperado desde la perspectiva clínica: los pacientes con la Enfermedad de Alzheimer presentan la edad promedio más elevada, de 74.74 años, seguidos por el grupo MCI con 72.21 años y, finalmente, el grupo CN con 70.51 años. Esta progresión es consistente con el hecho de que la edad es el principal factor de riesgo para el desarrollo de la enfermedad.

El Cuadro 6.3 presenta la distribución detallada por grupo diagnóstico y género. En el grupo AD, se observa una mayor proporción de hombres (325 sujetos, 55.46 %) en comparación con las mujeres (261 sujetos, 44.54 %), con edades promedio de 75.64 años y 73.62 años, respectivamente. Por el contrario, en el grupo CN predominan las mujeres (1,109 sujetos, 59.95 %) sobre los hombres (741 sujetos, 40.05 %), con edades promedio de 69.65 años y 71.79 años, respectivamente. En el grupo MCI, la distribución por género es más equilibrada, con 1,209 hombres (55.43 %) y 972 mujeres (44.57 %), con edades promedio de 73.31 años y 70.83 años, respectivamente.

Cuadro 6.3: Distribución por grupo diagnóstico y género

Grupo	**Género**	**Cantidad**	**% del Grupo**	**Edad Promedio**	**Desv. Estándar**
AD	Hombre	325	55.46 %	75.64 años	7.79
AD	Mujer	261	44.54 %	73.62 años	8.34
CN	Hombre	741	40.05 %	71.79 años	6.78
CN	Mujer	1,109	59.95 %	69.65 años	7.17
MCI	Hombre	1,209	55.43 %	73.31 años	7.31
MCI	Mujer	972	44.57 %	70.83 años	7.73

6.1.2.2. Descripción de las variables clínicas y sociodemográficas

El Cuadro 6.6 presenta un resumen de todas las variables clínicas y sociodemográficas utilizadas en este estudio, incluyendo su tipo, unidad de medición o codificación y su relación con la enfermedad de Alzheimer.

Cuadro 6.4: Resumen de variables clínicas y sociodemográficas

Variable	**Tipo**	**Unidad / Codificación**	**Relación con la EA**
MMSE Score	Numérica continua	0–30 puntos	A menor puntuación, mayor deterioro cognitivo
Edad	Numérica continua	Años	Mayor edad, mayor riesgo de EA
Género	Categórica	1 = Hombre, 2 = Mujer	Mayor prevalencia en mujeres
Años de educación	Numérica discreta	Años	Más años, mayor reserva cognitiva
Grupo diagnóstico	Categórica nominal	CN, MCI, AD	Variable objetivo para clasificación
Genotipo APOE	Categórica ordinal	2/2, 2/3, ..., 4/4	Mayor riesgo con presencia del alelo *E4*
Peso	Numérica continua	Kilogramos	Relacionado con salud metabólica y vascular
Cardiopatías	Binaria	1 = Sí, 0 = No	Asociadas a mayor riesgo de demencia
Consumo de alcohol	Binaria	1 = Sí, 0 = No	Puede afectar la función cognitiva
Consumo de drogas	Binaria	1 = Sí, 0 = No	Factor de riesgo neurológico
Tabaquismo	Binaria	1 = Sí, 0 = No	Relación con daño vascular y cognitivo

6.2. Preprocesamiento

En este apartado se describen las principales decisiones metodológicas adoptadas durante las etapas de preprocesamiento de los datos. Dichas decisiones incluyen la división de la base de datos, así como los procedimientos de limpieza, transformación y reorganización de la información, los cuales fueron necesarios para garantizar la calidad, coherencia y adecuación de los datos de entrada antes de su utilización en las fases posteriores de análisis y modelado.

6.2.1. Preprocesamiento de los estudios PET

6.2.1.1. División del dataset

Para garantizar la evaluación rigurosa del modelo y evitar el sobreajuste, el conjunto de datos completo fue dividido en tres conjuntos independientes: entrenamiento (*train*), validación (*val*) y prueba (*test*), siguiendo una estrategia de división estratificada a nivel de estudio.

1. Estrategia de División:

La división del dataset se realizó mediante un proceso aleatorio controlado con una semilla fija (`RANDOM_SEED = 42`) para garantizar la reproducibilidad de los resultados. En donde se establecieron las siguientes proporciones:

- **Conjunto de entrenamiento**: 80 % de los estudios
- **Conjunto de validación**: 18 % de los estudios
- **Conjunto de prueba**: 2 % de los estudios

La división se implementó preservando la estructura jerárquica del dataset, asegurando que se asignaran estudios completos y no imágenes individuales a cada conjunto. Este enfoque metodológico previene la fuga de información entre conjuntos, situación que ocurriría si múltiples cortes del mismo estudio se distribuyeran entre diferentes splits.

Ahora bien, el proceso de división se llevo a cabo a nivel de carpeta de estudio, donde cada carpeta contenía archivos en formato DICOM o NIfTI correspondientes a un único examen PET. La asignación aleatoria se realizó de manera independiente para cada combinación de grupo diagnóstico y categoría patológica, manteniendo la representatividad de todas las clases en los tres conjuntos.

2. Resultados de la división:

El Cuadro 6.5 presenta la distribución final de los estudios en cada conjunto:

Cuadro 6.5: Distribución de estudios en los conjuntos de entrenamiento, validación y prueba

Conjunto	N° de estudios	Porcentaje
Entrenamiento	4,535	79.9 %
Validación	1,018	17.9 %
Prueba	120	2.1 %
Total	**5,673**	**100 %**

La distribución obtenida se aproxima a las proporciones establecidas, con ligeras variaciones debido al redondeo inherente al número discreto de estudios por clase. El conjunto de entrenamiento, con 4,535 estudios, proporciona suficientes datos para el aprendizaje de patrones complejos. Por otro lado, el conjunto de validación, con 1,018 estudios, permite el ajuste de hiperparámetros y la selección del modelo óptimo durante el entrenamiento. Finalmente, el conjunto de prueba, aunque reducido, con 120 estudios, mantiene representación de todas las categorías diagnósticas para la evaluación final del modelo.

6.2.1.2. Conversión de formato DICOM a NIfTI

Las imágenes PET originales de ADNI se encuentran almacenadas en formato DICOM (Digital Imaging and Communications in Medicine), que es el estándar clínico para el almacenamiento y transmisión de imágenes médicas. Sin embargo, para facilitar el procesamiento computacional y la manipulación matricial de los datos volumétricos, fue necesario convertir todas las imágenes al formato NIfTI (Neuroimaging Informatics Technology Initiative).

Esta conversión se realizó mediante la herramienta dcm2niix [89], un conversor optimizado que preserva toda la información geométrica y de metadatos de las imágenes DICOM originales. La herramienta extrae automáticamente los parámetros de espaciado de voxel, orientación espacial y origen del volumen desde los encabezados DICOM, transfiriéndolos al archivo NIfTI resultante. Este proceso se ejecutó de forma automatizada para cada uno de los estudios PET del dataset, generando archivos NIfTI comprimidos (.nii.gz) que facilitan tanto el almacenamiento como la carga eficiente de datos.

Durante esta etapa, se implementaron verificaciones de integridad para asegurar que la conversión fuera exitosa. En casos donde las imágenes originales presentaban dimensiones no estándar, por ejemplo, volúmenes 4D con una dimensión temporal de un solo punto, se realizó una extracción automática del primer volumen temporal para garantizar que todas las imágenes resultantes fueran tridimensionales. Adicionalmente, se detectaron y corrigieron casos excepcionales de imágenes con valores de intensidad extremadamente bajos o nulos, los cuales podrían interferir con las etapas posteriores de registro, mediante la adición de un mínimo de ruido gaussiano que preserva la estructura de la imagen mientras previene fallos numéricos.

6.2.1.3. Normalización espacial mediante registro de imágenes

Una de las principales fuentes de variabilidad en el dataset es la orientación espacial y la posición de la cabeza del paciente durante la adquisición. Para eliminar esta variabilidad y garantizar que todas las estructuras anatómicas estén alineadas en un espacio común, se implementó un procedimiento de registro rígido de imágenes utilizando un atlas PET como referencia.

El atlas de referencia seleccionado corresponde a una plantilla PET cerebral construida a partir de sujetos cognitivamente normales (CN_PET_TEMPLATE.nii), que representa la anatomía cerebral promedio de individuos sin deterioro cognitivo. Este atlas fue generado específicamente para estudios PET de la enfermedad de Alzheimer y captura las características de intensidad y distribución espacial típicas de las imágenes PET cerebrales.

El proceso de registro se realizó mediante transformaciones rígidas de seis grados de libertad (tres traslaciones y tres rotaciones), las cuales permiten alinear las imágenes sin introducir deformaciones no lineales que podrían alterar las características morfométricas del cerebro. La inicialización del registro se basó en el método de momentos de primer y segundo orden, que calcula una alineación inicial aproximada mediante el centrado de las imágenes y la alineación de sus ejes principales de inercia. Esta inicialización robusta es particularmente importante para garantizar la convergencia del algoritmo de optimización posterior.

La métrica de similitud utilizada fue la Información Mutua de Mattes (Mattes Mutual Infor-

mation) [90], una métrica ampliamente utilizada en el registro multimodal que no asume ninguna relación funcional específica entre las intensidades de las imágenes a registrar. Esta métrica es particularmente apropiada para PET, donde las relaciones de intensidad pueden variar significativamente entre diferentes radiofármacos. Se utilizaron 50 bins de histograma para la estimación de la información mutua, proporcionando un balance adecuado entre precisión y eficiencia computacional.

La optimización de la transformación se llevó a cabo mediante el gradiente descendente, con un tamaño de paso inicial de 1.0 y un máximo de 100 iteraciones. Durante este proceso, el algoritmo ajusta iterativamente los parámetros de la transformación rígida para maximizar la información mutua entre la imagen a registrar y el atlas de referencia. Se utilizó interpolación lineal durante el proceso de registro para balancear la precisión con la eficiencia computacional.

En casos excepcionales en los que el registro rígido no convergió adecuadamente debido a características de imagen poco comunes o artefactos, se implementó un mecanismo de contingencia que aplica directamente un remuestreo al espacio del atlas sin optimización iterativa, garantizando así que todas las imágenes sean procesadas exitosamente.

6.2.1.4. Remuestreo y homogeneización espacial

Tras el registro rígido, todas las imágenes fueron remuestreadas al espacio geométrico del atlas de referencia. Este proceso garantiza que todas las imágenes del dataset compartan exactamente las mismas dimensiones matriciales, resolución espacial y orientación, lo cual es un requisito fundamental para el entrenamiento de redes neuronales convolucionales que esperan entradas de tamaño fijo.

El remuestreo se realizó mediante interpolación B-Spline de tercer orden, la cual proporciona una reconstrucción suave de las intensidades en las nuevas posiciones de voxel, mientras minimiza los artefactos de aliasing y preserva las transiciones graduales de intensidad características de las imágenes PET. La interpolación B-Spline es superior a la interpolación lineal para las imágenes médicas, ya que produce resultados más precisos en la reconstrucción de señales suaves y reduce los artefactos visuales.

Como resultado de esta etapa, todas las imágenes quedaron estandarizadas a las dimensiones del atlas de referencia, con un espaciado de voxel uniforme que facilita la comparación directa entre estudios y elimina la necesidad de consideraciones especiales sobre la resolución espacial durante el entrenamiento de los modelos.

6.2.1.5. Normalización de intensidades

Las imágenes PET presentan una variabilidad significativa en sus rangos de intensidad debido a diferencias en la dosis de radiofármaco administrada, el tiempo transcurrido entre la inyección y la adquisición, la sensibilidad del escáner y los factores de calibración específicos de cada centro. Para eliminar esta variabilidad y facilitar el aprendizaje de los modelos, se implementó un procedimiento de normalización de intensidades.

La estrategia de normalización adoptada se basa en percentiles robustos en lugar de valores extremos absolutos, lo cual proporciona mayor resistencia frente a valores atípicos o artefactos

puntuales. Específicamente, se calcularon los percentiles 1 y 99 de la distribución de intensidades de cada imagen, y todos los valores fueron recortados (clipping) a este rango antes de la normalización. Este procedimiento elimina el impacto de valores extremos que podrían deberse a ruido, artefactos de reconstrucción o regiones extracerebrales con alta captación.

Posteriormente, las intensidades recortadas fueron normalizadas linealmente al rango [0, 1], lo cual garantiza que todas las imágenes presenten rangos de intensidad comparables, facilitando la convergencia durante el entrenamiento de las redes neuronales y permitiendo el uso de tasas de aprendizaje consistentes a través de diferentes lotes de datos.

Previo a la normalización, se aplicó una conversión de valores NaN (Not a Number) a cero para garantizar la estabilidad numérica del procesamiento. Estos valores NaN pueden aparecer ocasionalmente en imágenes PET debido a errores en la reconstrucción o en regiones sin información.

6.2.1.6. Aplicación de máscara cerebral

Como etapa final del pipeline de preprocesamiento, se aplicó una máscara cerebral binaria para aislar el tejido cerebral y eliminar estructuras extracerebrales que no son relevantes para el análisis de la enfermedad de Alzheimer. Esta máscara (CN_PET_TEMPLATE_MASK.nii) fue diseñada específicamente para el atlas de referencia utilizado y delimita con precisión los contornos del parénquima cerebral.

La aplicación de la máscara cerebral cumple varios propósitos importantes. Primero, elimina regiones como el cuero cabelludo, el hueso craneal y los tejidos blandos extracraneales que presentan captación variable del radiofármaco, pero no aportan información diagnóstica. Segundo, reduce significativamente el espacio de características que los modelos deben procesar, mejorando la eficiencia computacional. Tercero, ayuda a enfocar el aprendizaje de las redes neuronales en las regiones anatómicamente relevantes, lo que potencialmente mejora su capacidad de generalización.

La máscara fue remuestreada al espacio de cada imagen normalizada utilizando interpolación de vecino más cercano (nearest neighbor), que es la técnica apropiada para imágenes binarias, ya que preserva los valores discretos sin introducir valores intermedios. Tras el remuestreo, la máscara fue convertida a tipo entero sin signo de 8 bits y aplicada mediante una operación de enmascaramiento que establece en cero todos los voxeles extracerebrales, mientras preserva los valores normalizados originales para los voxeles intracerebrales.

6.2.1.7. Organización y almacenamiento de datos preprocesados

Las imágenes resultantes del pipeline de preprocesamiento fueron organizadas manteniendo la estructura jerárquica original del dataset: división (train/val/test), grupo funcional (Amiloide/Metabólico/Tau) y categoría diagnóstica (AD/CN/MCI). Esta organización facilita el acceso eficiente a los datos durante el entrenamiento y la evaluación de los modelos.

6.2.2. Preprocesamiento de los datos sociodemográficos

6.2.2.1. Normalización y limpieza de datos

Inicialmente, se llevó a cabo una etapa de preprocesamiento orientada a la estandarización de los nombres de las columnas del conjunto de datos. Este procedimiento consistió en la eliminación de espacios en blanco redundantes y en la conversión de todos los identificadores a letras mayúsculas, con el propósito de unificar la nomenclatura de las variables. Esta normalización resulta fundamental para garantizar la consistencia en el acceso a los atributos del conjunto de datos, minimizar ambigüedades semánticas y prevenir errores derivados de inconsistencias en la denominación de las columnas durante las etapas posteriores de análisis y modelado.

Posteriormente, se realizó la transformación del formato de las variables numéricas que originalmente empleaban la convención europea, caracterizada por el uso de la coma como separador decimal, al formato estándar anglosajón, que utiliza el punto como separador decimal. Esta conversión se aplicó específicamente a las variables *MMSE Score*, *Age*, *Weight* y *Education*. Para ello, se emplearon expresiones regulares que permitieron realizar el reemplazo sistemático del separador decimal, seguido de la conversión explícita de los valores a tipo numérico mediante la función `pd.to_numeric()`. Adicionalmente, se implementó un manejo robusto de errores a través del parámetro `errors='coerce'`, el cual fuerza la conversión de valores no interpretables a `NaN`, facilitando así su posterior identificación y tratamiento dentro del flujo de limpieza y preparación de los datos.

6.2.2.2. Codificación y transformación de variables categóricas

Las variables categóricas requirieron la aplicación de transformaciones específicas para garantizar su correcta integración en los modelos de aprendizaje automático. Dado que los algoritmos empleados operan exclusivamente sobre representaciones numéricas, fue indispensable convertir dichas variables en formatos cuantificables mediante esquemas de codificación adecuados. Este proceso permitió preservar, en la medida de lo posible, la información semántica inherente a cada categoría, evitando la introducción de sesgos artificiales en el modelo.

En particular, se seleccionaron técnicas de codificación acordes a la naturaleza de cada variable y a los supuestos de los algoritmos utilizados, de modo que se mantuviera la interpretabilidad y se optimizara el desempeño predictivo.

1. Procesamiento del genotipo APOE:

La variable correspondiente al genotipo *APOE* fue sometida a un proceso de preprocesamiento orientado a la extracción sistemática de patrones numéricos a partir de su representación textual. Para este fin, se emplearon expresiones regulares, específicamente el patrón `r'(\d\/\d)'`, el cual permitió identificar la notación estándar de los alelos, tales como 3/3, 3/4 y 4/4, presentes en el conjunto de datos.

Con base en la información extraída, se definieron dos variables derivadas que facilitan una representación más informativa del riesgo genético asociado al genotipo *APOE*. En primer lugar, se construyó la variable **APOE_E4**, correspondiente a un indicador binario (0/1) que señala la presencia del alelo $\varepsilon 4$, ampliamente reconocido como el principal factor de riesgo genético para el

desarrollo de la enfermedad de Alzheimer de inicio tardío [91]. Esta variable fue generada mediante una función lambda que evalúa la aparición del carácter `'4'` en la cadena asociada al genotipo.

Adicionalmente, se incorporó la variable **APOE_MISSING**, también de naturaleza binaria, cuyo objetivo es identificar la ausencia de información genotípica. La inclusión de este indicador resulta fundamental para permitir que los modelos de aprendizaje automático distingan explícitamente entre la ausencia del alelo de riesgo y la falta de datos genéticos, evitando así interpretaciones erróneas en las que los valores faltantes podrían ser tratados implícitamente como ausencia de riesgo.

2. Codificación de género:

La variable correspondiente al género biológico fue codificada numéricamente siguiendo la convención establecida por la *Alzheimer's Disease Neuroimaging Initiative* (ADNI) [66]. En el conjunto de datos original, esta variable se encontraba representada mediante valores enteros, donde el valor 1 correspondía al género masculino y el valor 2 al género femenino. Con el fin de estandarizar su representación para su uso en modelos de aprendizaje automático, se aplicó un proceso de remapeo que asigna el valor 0 al género masculino y el valor 1 al género femenino.

Dicha transformación se implementó utilizando el método `map()` de la librería `pandas`, mediante un diccionario de correspondencias definido como `{1: 0, 2: 1}`. Este enfoque permitió una conversión directa y transparente de la variable, asegurando la integridad de la información original y evitando ambigüedades en la interpretación de los datos.

3. Codificación de la variable objetivo:

La variable objetivo, correspondiente al diagnóstico clínico, constituye la variable dependiente que los modelos de aprendizaje automático buscan predecir. Dado que esta variable se encontraba originalmente representada mediante categorías textuales heterogéneas, fue necesario implementar un proceso de transformación orientado a su conversión en clases numéricas ordenadas. Este procedimiento facilita tanto el procesamiento computacional por parte de los algoritmos de clasificación como la interpretación clínica de los resultados obtenidos.

El esquema de codificación adoptado se definió en función de los principales estados clínicos asociados al espectro de la enfermedad de Alzheimer, de acuerdo con la progresión cognitiva reportada en la literatura. En consecuencia, se establecieron las siguientes clases:

- **Clase 0 = Control Normal (CN)**: Individuos cognitivamente sanos, sin evidencia clínica de deterioro cognitivo.
- **Clase 1 = Deterioro Cognitivo Leve (MCI)**: Pacientes que presentan déficits cognitivos detectables mediante evaluación clínica o neuropsicológica, los cuales no interfieren de manera significativa con las actividades de la vida diaria.
- **Clase 2 = Enfermedad de Alzheimer (AD)**: Pacientes con diagnóstico clínico confirmado de demencia tipo Alzheimer.

La implementación de este mapeo se realizó mediante una función personalizada diseñada para evaluar la presencia de palabras clave específicas dentro de las etiquetas diagnósticas originales. Como etapa previa, las cadenas de texto fueron normalizadas mediante la conversión a mayúsculas y la eliminación de espacios en blanco utilizando el método `str.upper().str.strip()`, con el

fin de reducir la variabilidad semántica y garantizar una clasificación consistente. Posteriormente, se aplicaron evaluaciones condicionales para identificar términos representativos de cada categoría diagnóstica, tales como `'CN'`, `'CONTROL'` y `'NORMAL'` para la clase 0; `'MCI'`, `'MILD'` y `'DCL'` para la clase 1; y `'AD'` y `'ALZHEIMER'` para la clase 2.

6.2.2.3. Selección y clasificación de características

El conjunto final de características se seleccionó basándose en su relevancia clínica documentada en la literatura sobre la enfermedad de Alzheimer y su disponibilidad en el conjunto de datos. Las 11 características finales se clasificaron en dos categorías funcionalmente distintas para su procesamiento diferenciado, las cuales se contienen en el Cuadro 6.6:

Cuadro 6.6: Características clínicas y demográficas utilizadas en el modelo

Tipo de característica	**Variable**	**Descripción**
Numéricas continuas	MSESCORE	Puntaje del Mini-Mental State Examination (MMSE).
	AGE	Edad del paciente (años).
	EDUCATION	Nivel educativo del paciente.
	WEIGHT	Peso corporal del paciente.
Binarias categóricas (0/1)	GENDER_CODE	Sexo del paciente.
	APOE_E4	Presencia del alelo APOE $\varepsilon 4$.
	APOE_MISSING	Indicador de dato faltante para APOE.
	CARDIOPATHY	Antecedentes de enfermedad cardiovascular.
	ALCOHOLISM	Antecedentes de alcoholismo.
	DRUGS	Consumo de sustancias psicoactivas.
	SMOK	Hábito de fumar.

6.2.3. Transformación de características

Para las características numéricas, se implementó un pipeline secuencial de transformaciones utilizando la clase `Pipeline` de scikit-learn. Este pipeline garantiza que todas las transformaciones se apliquen en el orden correcto y de manera reproducible tanto en el conjunto de entrenamiento como en el de prueba.

6.2.3.1. Imputación de valores faltantes

La presencia de valores faltantes en conjuntos de datos clínicos es una situación frecuente, derivada de pruebas no realizadas, información incompleta en registros históricos o pérdidas durante

el proceso de recolección. Para abordar esta problemática, se utilizó la clase `SimpleImputer` de `scikit-learn`, empleando una estrategia de imputación basada en la mediana.

La selección de la mediana como estadístico de imputación se justifica por su mayor robustez frente a valores atípicos en comparación con la media aritmética [92]. Esta característica resulta especialmente relevante en el contexto clínico, donde la presencia de mediciones extremas puede reflejar condiciones reales de ciertos pacientes y no debería influir de manera desproporcionada en el proceso de imputación.

La imputación se realizó de forma independiente para cada variable numérica, calculando la mediana únicamente sobre los valores observados del conjunto de entrenamiento y aplicando posteriormente este valor para reemplazar los valores faltantes (`NaN`) tanto en los datos de entrenamiento como en los de prueba.

6.2.3.2. Estandarización de características

Posterior al proceso de imputación, se aplicó una estandarización de las variables numéricas utilizando la clase `StandardScaler` de `scikit-learn`. Esta transformación normaliza cada característica para que presente media cero ($\mu = 0$) y desviación estándar unitaria ($\sigma = 1$), de acuerdo con la expresión:

$$z = \frac{x - \mu}{\sigma} \tag{6.1}$$

donde x corresponde al valor original de la característica, mientras que μ y σ representan la media y la desviación estándar calculadas sobre el conjunto de entrenamiento, respectivamente.

La estandarización resulta especialmente relevante para algoritmos sensibles a la escala de las variables, como *K-Nearest Neighbors* (KNN), en los cuales la distancia euclidiana determina la asignación de clase [93]. En ausencia de esta transformación, variables con un mayor rango numérico pueden dominar el cálculo de distancias y sesgar las predicciones.

Si bien modelos como *Random Forest* y *Naive Bayes* presentan menor sensibilidad a la escala, la estandarización contribuye a una mejor estabilidad numérica y facilita la interpretación comparativa de las características. Es importante resaltar que los parámetros de estandarización (μ y σ) se estimaron exclusivamente a partir del conjunto de entrenamiento y, posteriormente, se aplicaron al conjunto de prueba, evitando así la fuga de información (*data leakage*) [94].

6.2.3.3. Tratamiento de características binarias

Las características de naturaleza binaria no requirieron transformaciones adicionales más allá de la codificación numérica previamente aplicada. En consecuencia, se adoptó una estrategia de *passthrough* (paso directo), mediante la cual se preservaron los valores originales 0/1 sin modificación.

Esta decisión metodológica se sustenta en varios aspectos. En primer lugar, las variables binarias ya se encuentran en una escala homogénea y comparable. En segundo lugar, la aplicación de técnicas de estandarización sobre este tipo de variables podría comprometer su interpretabilidad semántica al alterar el significado lógico asociado a cada categoría. Finalmente, los algoritmos de aprendizaje

automático empleados en este estudio manejan de forma eficiente características binarias sin requerir transformaciones adicionales, lo que hace innecesario cualquier ajuste posterior.

6.2.3.4. Implementación con columnTransformer

La diferenciación entre las transformaciones aplicadas a las características numéricas y binarias se implementó mediante el uso de la clase `ColumnTransformer` de `scikit-learn` [95]. Esta herramienta permite aplicar transformadores específicos a subconjuntos definidos de columnas, integrándolos en un único transformador compuesto que opera de forma consistente sobre todo el *DataFrame*.

En la configuración adoptada se especificaron dos transformadores principales. El primero, denominado `'num'`, corresponde a un *pipeline* que combina la imputación de valores faltantes y la estandarización, aplicado exclusivamente a las características numéricas. El segundo, `'bin'`, implementa una estrategia de *passthrough* para las características binarias, preservando sus valores originales sin modificación. Asimismo, el parámetro `remainder='drop'` garantiza que cualquier columna no definida explícitamente sea excluida del conjunto de datos transformado, proporcionando un control estricto sobre las variables utilizadas en el modelado.

La arquitectura de `ColumnTransformer` asegura que el orden de las columnas en la salida sea determinístico y reproducible, un aspecto crítico para la consistencia del *pipeline* de aprendizaje automático. Adicionalmente, su integración nativa con la clase `Pipeline` permite construir flujos completos de preprocesamiento y modelado que se comportan como una única entidad, facilitando el uso estandarizado de los métodos `fit()`, `transform()` y `predict()` a lo largo de todo el proceso experimental.

6.2.4. División del conjunto de datos

La partición del conjunto de datos se realizó de acuerdo con las buenas prácticas ampliamente recomendadas en la literatura de aprendizaje automático [96]. En particular, se implementó una estrategia de tipo *hold-out* o *train-test split*, asignando el 80 % de las muestras al conjunto de entrenamiento y el 20 % restante al conjunto de prueba. Esta división se llevó a cabo mediante la función `train_test_split` de `scikit-learn`, garantizando una separación clara entre los datos utilizados para el ajuste del modelo y aquellos empleados para su evaluación.

El Cuadro 6.7 resume la distribución final de las muestras en ambos conjuntos, evidenciando una proporción equilibrada que permite un entrenamiento adecuado del modelo sin comprometer la representatividad del conjunto de prueba.

Cuadro 6.7: Distribución de muestras en conjuntos de entrenamiento y prueba

Conjunto	**Número de Muestras**	**Porcentaje**
Entrenamiento	3,693	80.0 %
Prueba	924	20.0 %
Total	**4,617**	**100.0 %**

6.2.4.1. Estratificación de la división

La estratificación de la variable objetivo, habilitada mediante el parámetro `stratify=y`, constituye un componente fundamental en el proceso de partición de los datos. Este procedimiento garantiza que la distribución de las clases diagnósticas en los conjuntos de entrenamiento y prueba reproduzca de forma proporcional la distribución observada en el conjunto de datos original [97].

En el contexto de este estudio, caracterizado por un desbalance significativo entre clases (40.1 % CN, 47.2 % MCI y 12.7 % AD), la estratificación previene escenarios indeseables, como la subrepresentación de la clase minoritaria, que es la enfermedad de Alzheimer; la generación de particiones no representativas debido a la variabilidad aleatoria; o la evaluación del desempeño del modelo sobre bases muestrales inadecuadas que podrían sesgar las métricas de evaluación.

El Cuadro 6.8 presenta la distribución de las clases diagnósticas en el conjunto original y demuestra la preservación efectiva de dichas proporciones en los conjuntos de entrenamiento y prueba tras aplicar la estrategia de estratificación. Como se observa, las diferencias porcentuales entre los conjuntos son mínimas y atribuibles únicamente al redondeo inherente a la partición de un número discreto de muestras. Esta consistencia garantiza que el conjunto de prueba constituya una muestra representativa de la población objetivo, permitiendo comparaciones válidas y una evaluación robusta de los modelos de clasificación.

Cuadro 6.8: Distribución estratificada de clases diagnósticas en los conjuntos de datos

Clase	**Dataset Original**		**Entrenamiento**		**Prueba**	
	n	**%**	**n**	**%**	**n**	**%**
Control Normal (CN)	1,850	40.1	1,480	40.1	370	40.0
MCI	2,181	47.2	1,744	47.2	437	47.3
Alzheimer (AD)	586	12.7	469	12.7	117	12.7
Total	**4,617**	**100.0**	**3,693**	**100.0**	**924**	**100.0**

6.2.4.2. Validación cruzada

De manera complementaria a la partición del dataset *train-test*, se implementó un esquema de validación cruzada estratificada de cinco pliegues (*5-fold stratified cross-validation*) sobre el conjunto de entrenamiento. Esta técnica divide iterativamente los datos en cinco subconjuntos, utilizando cuatro para el entrenamiento y uno para la validación en cada iteración, rotando de forma sistemática el subconjunto de validación [97].

La validación cruzada estratificada permite obtener estimaciones más robustas y estables del desempeño del modelo en comparación con una única división de *train-test*, al reducir la varianza asociada a una selección específica de datos de validación. Asimismo, posibilita evaluar la consistencia del modelo mediante la variabilidad de las métricas entre pliegues y facilita la detección de posibles escenarios de sobreajuste cuando se observan discrepancias significativas entre el rendimiento en validación y en el conjunto de prueba final.

En este estudio, este esquema de validación cruzada se utilizó específicamente durante el proceso de optimización de hiperparámetros mediante la clase `GridSearchCV`, asegurando que la selección de

las configuraciones óptimas se basara en estimaciones confiables del rendimiento de generalización.

6.3. Modelos implementados

En el presente estudio, se implementaron dos categorías principales de modelos de aprendizaje automático, diferenciadas según la naturaleza de los datos de entrada. Por un lado, se emplearon modelos de aprendizaje profundo basados en redes neuronales convolucionales tridimensionales (3D-CNN) para el procesamiento de imágenes PET volumétricas. Por otro lado, se utilizaron modelos de aprendizaje automático clásico (*Machine Learning*, ML) orientados al análisis de datos tabulares de carácter sociodemográfico y clínico.

El Cuadro 6.9 presenta un resumen comparativo de los modelos implementados en este trabajo, categorizados de acuerdo con el tipo de datos que procesan, el paradigma de aprendizaje empleado y el número de variantes evaluadas para cada arquitectura.

Cuadro 6.9: Resumen de modelos implementados según tipo de datos de entrada

Tipo de Datos	**Paradigma**	**Modelo**	**Variantes**
Imágenes PET 3D	Deep Learning	ResNet3D	1
		Transfer Learning ResNet10-3D	1
		VoxCNN3D	1
Datos Tabulares	Machine Learning	K-Nearest Neighbors	2
		Naive Bayes Gaussiano	1
		Random Forest	1
Total de configuraciones evaluadas			**7**

6.3.1. Modelos de Redes Neuronales Convolucionales (CNN)

Para el procesamiento y la clasificación de los estudios PET, se implementaron tres arquitecturas de redes neuronales convolucionales diseñadas específicamente para el análisis de datos volumétricos. La selección de estas arquitecturas se basó en su capacidad para aprender representaciones jerárquicas de características espaciales tridimensionales, su eficiencia computacional y su desempeño previamente reportado en la literatura para tareas de clasificación de neuroimágenes médicas. Estas propiedades las convierten en herramientas adecuadas para la identificación de patrones metabólicos asociados a los diferentes estadios tempranos de la enfermedad de Alzheimer.

6.3.1.1. Modelo ResNet3D

Se implementó una arquitectura *ResNet3D* personalizada, inspirada en los principios de las redes residuales propuestos por He et al. [98], y adaptada específicamente para operar sobre volúmenes tridimensionales. Este tipo de arquitectura introduce conexiones residuales (*skip connections*) que permiten el flujo directo de la información y de los gradientes a través de capas profundas.

El uso de estas conexiones mitiga el problema de degradación del gradiente asociado al incremento en la profundidad de la red, facilitando un entrenamiento más estable y eficiente. En consecuencia, la arquitectura *ResNet3D* resulta adecuada para la extracción de representaciones jerárquicas complejas a partir de imágenes PET volumétricas, donde la preservación de información espacial tridimensional es fundamental para la tarea de clasificación.

La arquitectura implementada consta de los siguientes componentes estructurales, presentados en el Cuadro 6.10:

Cuadro 6.10: Arquitectura del modelo ResNet3D implementado

Componente	**Operación**	**Filtros**	**Kernel**	**Stride**
Stem	Conv3D	32	$7 \times 7 \times 7$	2
	BatchNorm + ReLU	-	-	-
	MaxPool3D	-	$3 \times 3 \times 3$	2
Bloque Residual 1	ResBlock	32	$3 \times 3 \times 3$	1
Bloque Residual 2	ResBlock	64	$3 \times 3 \times 3$	2
Bloque Residual 3	ResBlock	128	$3 \times 3 \times 3$	2
Bloque Residual 4	ResBlock	256	$3 \times 3 \times 3$	2
Clasificador	21	-	-	-
	Dropout (0.5)	-	-	-
	Dense + Softmax	3	-	-

Cada bloque residual implementa la siguiente estructura funcional:

1. Convolución 3D, seguida de normalización por lotes y activación ReLU
2. Segunda convolución 3D, seguida de normalización por lotes
3. Adición de la conexión residual (shortcut)
4. Activación ReLU final

La conexión residual se implementó mediante una proyección lineal en aquellos casos en los que las dimensiones espaciales o el número de canales difieren entre la entrada y la salida del bloque residual, garantizando así la compatibilidad dimensional necesaria para la operación de suma. Este mecanismo permite preservar la integridad de la información propagada a través de la red, independientemente de los cambios estructurales entre capas consecutivas.

Una característica distintiva de esta implementación es la aplicación conservadora de regularización. En particular, se incorporó únicamente una capa de *dropout* con una probabilidad de 0.5 antes de la capa de clasificación final, evitando la introducción de regularización excesiva dentro de los bloques residuales. Esta decisión busca prevenir la degradación de la capacidad de aprendizaje de representaciones espaciales complejas, aspecto especialmente relevante en tareas de clasificación de neuroimágenes tridimensionales [99].

6.3.1.2. Modelo Transfer Learning ResNet10-3D

El enfoque de *transfer learning* se implementó a partir de una arquitectura *ResNet-10* tridimensional preentrenada sobre el conjunto de datos *MedicalNet* [100], el cual integra múltiples modalidades de imágenes médicas volumétricas. El uso de pesos preentrenados permite reutilizar representaciones de bajo y medio nivel previamente aprendidas a partir de un corpus amplio y heterogéneo de datos médicos.

Esta estrategia resulta particularmente ventajosa cuando el conjunto de datos objetivo es de tamaño limitado, ya que acelera la convergencia durante el entrenamiento y puede contribuir a una mejora en el desempeño del modelo final. En este contexto, el *transfer learning* facilita una inicialización informada de los parámetros de la red, favoreciendo una mejor capacidad de generalización en tareas de clasificación de neuroimágenes PET tridimensionales [101].

El proceso de *transfer learning* se estructuró en dos fases claramente diferenciadas, con el objetivo de maximizar el aprovechamiento de las representaciones preentrenadas y asegurar un entrenamiento estable del modelo:

Fase 1: Entrenamiento de las capas superiores

En una primera etapa, se congeló por completo el *backbone* preentrenado, incluyendo todas las capas convolucionales y las capas de normalización por lotes. De este modo, únicamente se permitió la actualización de los parámetros correspondientes a las nuevas capas superiores diseñadas específicamente para la tarea de clasificación multiclase. Esta estrategia evita la degradación prematura de las representaciones aprendidas durante el preentrenamiento, particularmente en las fases iniciales del ajuste, cuando las capas añadidas aún no han alcanzado la convergencia [73].

Las capas superiores implementadas consistieron en una secuencia de capas densas con funciones de activación ReLU y regularización mediante *dropout*, tal como se resume en el Cuadro 6.11. Esta configuración busca equilibrar la capacidad de discriminación del modelo con la prevención del sobreajuste, favoreciendo una adecuada generalización.

Cuadro 6.11: Arquitectura de las capas superiores del modelo Transfer Learning

Capa	**Entrada**	**Salida**	**Dropout**
Dense 1	512	256	0.5
Dense 2	256	128	0.4
Dense 3 (salida)	128	3	-

Fase 2: Ajuste fino completo (*Fine-tuning*)

Una vez alcanzada la convergencia de las capas superiores, se procedió al descongelamiento completo del *backbone* preentrenado, permitiendo la actualización conjunta de todos los parámetros del modelo mediante un proceso de *fine-tuning*. En esta etapa, se empleó una tasa de aprendizaje reducida, aproximadamente un orden de magnitud inferior a la utilizada en la fase inicial, con el objetivo de realizar ajustes graduales sobre las representaciones preentrenadas sin alterar de manera significativa el conocimiento previamente adquirido [102].

Este enfoque permite adaptar de forma controlada las características aprendidas durante el

preentrenamiento a las particularidades del conjunto de datos objetivo, favoreciendo una mejor especialización del modelo en la tarea de clasificación. El Cuadro 6.12 presenta un resumen de la distribución de parámetros del modelo completo tras el proceso de descongelamiento.

Cuadro 6.12: Distribución de parámetros en el modelo Transfer Learning ResNet10-3D

Componente	**Parámetros**	**Porcentaje**
Backbone ResNet-10 (congelado Fase 1)	14,237,571	98.1 %
Capas superiores personalizadas	282,560	1.9 %
Total (Fase 2)	**14,520,131**	**100.0 %**

6.3.1.3. Modelo VoxCNN3D

El modelo *VoxCNN3D* corresponde a una arquitectura personalizada diseñada para lograr un equilibrio entre la capacidad representacional y la eficiencia computacional. Esta red adopta una estrategia de codificación jerárquica basada en bloques convolucionales secuenciales, en los cuales se incrementa progresivamente la profundidad de las características mientras se reducen de manera controlada las dimensiones espaciales del volumen de entrada.

La arquitectura se organiza en tres componentes principales:

1. Stem convolucional: Capa inicial de convolución tridimensional con un kernel de gran tamaño ($7 \times 7 \times 7$) y un *stride* de 2, cuya finalidad es capturar patrones de bajo nivel y realizar una reducción temprana de las dimensiones espaciales del volumen de entrada, disminuyendo así la carga computacional de las capas posteriores.

2. Bloques convolucionales: Tres bloques secuenciales que implementan una estructura homogénea orientada a la extracción progresiva de características de mayor nivel. En donde cada bloque incluye:

- Dos convoluciones 3D consecutivas, cada una seguida de normalización por lotes y una función de activación ReLU.
- Una operación de *max-pooling* tridimensional para la reducción de la resolución espacial.
- Una capa de *dropout* con probabilidades crecientes (0.15, 0.20 y 0.25) se aplica como mecanismo de regularización progresiva a medida que aumenta la complejidad de las representaciones.

3. Clasificador: Un módulo final compuesto por una operación de *global average pooling* 3D, seguido de una capa densa con regularización mediante *dropout* (0.3) y una capa de salida con una función de activación *softmax*, encargada de producir las probabilidades asociadas a cada una de las clases diagnósticas.

El Cuadro 6.13 presenta la configuración detallada de los bloques que conforman la arquitectura *VoxCNN3D*.

Cuadro 6.13: Arquitectura del modelo VoxCNN3D

Bloque	Operación	Filtros	Kernel	Pool	Dropout
Stem	Conv3D + BN + ReLU	24	$7 \times 7 \times 7$	-	-
Bloque 1	Conv3D + BN + ReLU	48	$3 \times 3 \times 3$	-	-
	Conv3D + BN + ReLU	48	$3 \times 3 \times 3$	-	-
	MaxPool3D	-	$2 \times 2 \times 2$	Sí	-
	Dropout	-	-	-	0.15
Bloque 2	Conv3D + BN + ReLU	96	$3 \times 3 \times 3$	-	-
	Conv3D + BN + ReLU	96	$3 \times 3 \times 3$	-	-
	MaxPool3D	-	$2 \times 2 \times 2$	Sí	-
	Dropout	-	-	-	0.20
Bloque 3	Conv3D + BN + ReLU	128	$3 \times 3 \times 3$	-	-
	Conv3D + BN + ReLU	128	$3 \times 3 \times 3$	-	-
	MaxPool3D	-	$2 \times 2 \times 2$	Sí	-
	Dropout	-	-	-	0.25
Clasificador	GlobalAvgPool3D	-	-	-	-
	Dense + ReLU	192	-	-	-
	Dropout	-	-	-	0.30
	Dense + Softmax	3	-	-	-

Esta arquitectura fue diseñada específicamente para evitar el colapso en la predicción hacia una única clase, fenómeno observado frecuentemente en arquitecturas con regularización excesiva aplicada a conjuntos de datos desbalanceados [103]. La tasa de aprendizaje se estableció en 3×10^{-4}, superior a la empleada en ResNet3D, con el objetivo de facilitar la convergencia sin promover el sobreajuste.

6.3.2. Modelos de *Machine Learning* (ML)

Para el procesamiento y la clasificación de los datos tabulares de carácter sociodemográfico y clínico, se implementaron cuatro modelos de aprendizaje automático clásico. La selección de estos algoritmos se fundamentó en su eficiencia computacional, su grado de interpretabilidad y su desempeño ampliamente documentado en la literatura para tareas de clasificación médica basadas en datos estructurados. En conjunto, estos modelos permiten explorar diferentes supuestos estadísticos y estrategias de decisión, proporcionando una base sólida para la comparación de resultados y el análisis del comportamiento predictivo sobre variables clínicas relevantes.

6.3.2.1. Modelo K-Nearest Neighbors (KNN)

El algoritmo *K-Nearest Neighbors* (KNN) es un método de clasificación no paramétrico y basado en instancias que asigna una etiqueta de clase a una nueva muestra en función de las clases de sus k vecinos más cercanos en el espacio de características [104]. La cercanía entre muestras se determina a partir de una métrica de distancia definida previamente, lo que permite capturar relaciones complejas y no lineales entre variables, sin asumir una forma funcional explícita del modelo.

La simplicidad conceptual de KNN, junto con su capacidad para adaptarse a la distribución local de los datos, lo convierte en una línea base adecuada para tareas de clasificación médica, especialmente cuando se trabaja con conjuntos de datos de tamaño moderado y características clínicas heterogéneas.

La implementación del KNN requiere la definición de varios hiperparámetros críticos, los cuales influyen directamente en el sesgo, la varianza y el desempeño del clasificador:

- **Número de vecinos (k):** Controla el tamaño del vecindario considerado para la votación de clase. Valores pequeños de k tienden a capturar patrones locales con mayor sensibilidad al ruido, mientras que valores mayores favorecen decisiones más estables y suavizadas.
- **Métrica de distancia:** Determina la noción de similitud entre muestras, siendo comunes las distancias euclidiana, Manhattan y Minkowski, las cuales son especialmente relevantes en espacios de características clínicas multidimensionales.
- **Esquema de ponderación:** Puede ser uniforme, donde todos los vecinos contribuyen de igual forma, o ponderado por distancia, otorgando mayor influencia a los vecinos más cercanos al punto de consulta.

Dado que KNN es altamente sensible a la escala de las características, se aplicó un proceso de estandarización a todas las variables numéricas, conforme al procedimiento descrito en la sección de preprocesamiento. Este paso garantiza que las variables clínicas con diferentes rangos y unidades de medida contribuyan de manera equitativa al cálculo de distancias, evitando sesgos en la asignación de clases.

6.3.2.2. Modelo K-Nearest Neighbors (KNN) con SMOTE

Una limitación inherente al conjunto de datos sociodemográficos y clínicos es el desbalance significativo entre las clases diagnósticas, donde la categoría minoritaria correspondiente a pacientes con Enfermedad de Alzheimer representa únicamente el 12.7 % del total de las muestras. Este desequilibrio puede inducir un sesgo del clasificador hacia la clase mayoritaria, comprometiendo la sensibilidad del modelo y reduciendo su capacidad para detectar correctamente casos positivos de la enfermedad, lo cual resulta particularmente crítico en aplicaciones de diagnóstico temprano [105].

Con el fin de mitigar los efectos negativos del desbalanceo de clases, se implementó una variante del modelo K-Nearest Neighbors que incorpora la técnica SMOTE (*Synthetic Minority Oversampling Technique*) [106] como estrategia de balanceo supervisado. SMOTE genera nuevas muestras sintéticas de la clase minoritaria mediante interpolación lineal entre una instancia y sus k vecinos más cercanos pertenecientes a la misma clase, enriqueciendo el espacio de características sin recurrir a la simple duplicación de observaciones existentes, como ocurre en el sobremuestreo aleatorio.

La aplicación de SMOTE permite mejorar la representación de la clase minoritaria durante el entrenamiento del clasificador, favoreciendo la construcción de fronteras de decisión más equilibradas y reduciendo la probabilidad de sobreajuste a patrones específicos de la clase mayoritaria. Esta

estrategia resulta especialmente adecuada en combinación con KNN, dado que ambos métodos se basan explícitamente en relaciones de proximidad en el espacio de características.

El procedimiento de balanceo se aplicó exclusivamente al conjunto de entrenamiento, preservando la distribución original de clases en el conjunto de prueba. Esta decisión metodológica garantiza una evaluación realista del desempeño del modelo y evita una estimación optimista de las métricas de clasificación, asegurando que los resultados reflejen adecuadamente la capacidad de generalización del sistema en escenarios clínicos reales. El Cuadro 6.14 presenta la distribución de clases antes y después de la aplicación de SMOTE.

Cuadro 6.14: Distribución de clases antes y después de aplicar SMOTE

Clase	**Antes de SMOTE**		**Después de SMOTE**	
	n	%	**n**	%
Control Normal (CN)	1,480	40.1	1,744	33.3
MCI	1,744	47.2	1,744	33.3
Alzheimer (AD)	469	12.7	1,744	33.3
Total	**3,693**	**100.0**	**5,232**	**100.0**

La implementación de SMOTE se realizó con $k = 5$ vecinos para la generación de muestras sintéticas, un valor estándar que equilibra la diversidad y la coherencia en las muestras generadas [107].

6.3.2.3. Modelo Naive Bayes

El clasificador *Naive Bayes* es un modelo probabilístico generativo fundamentado en el teorema de Bayes, que asume independencia condicional entre las características dado el valor de la clase [108]. Aunque esta hipótesis de independencia rara vez se satisface de forma estricta en conjuntos de datos reales, especialmente en contextos clínicos donde las variables suelen presentar correlaciones fisiológicas y demográficas, el modelo ha demostrado un desempeño competitivo en múltiples aplicaciones prácticas, junto con ventajas computacionales significativas.

En este estudio se implementó la variante *Gaussian Naive Bayes*, adecuada para variables continuas, en la cual la verosimilitud de cada característica se modela mediante una distribución normal parametrizada por la media y la varianza específicas de cada clase. Asimismo, una ventaja destacada de *Naive Bayes* en el contexto clínico es su capacidad para proporcionar estimaciones probabilísticas explícitas por clase, lo que facilita la interpretación de la confianza asociada a cada predicción y permite su integración en sistemas de apoyo a la decisión médica. Además, su bajo costo computacional y rápida inferencia lo convierten en una alternativa adecuada como modelo base y como punto de comparación frente a clasificadores más complejos [109].

6.3.2.4. Modelo Random Forest

Random Forest es un método de *ensemble learning* basado en la construcción de múltiples árboles de decisión entrenados de manera independiente, cuyas predicciones se combinan mediante

votación mayoritaria en tareas de clasificación o promediado en problemas de regresión [110]. Este enfoque busca mejorar la capacidad de generalización del modelo al reducir la varianza inherente a los árboles de decisión individuales.

La diversidad entre los árboles que conforman el bosque se garantiza mediante dos mecanismos fundamentales de aleatorización:

1. **Bootstrap aggregating (*bagging*):** cada árbol se entrena utilizando una muestra aleatoria con reemplazo del conjunto de entrenamiento original, lo que introduce variabilidad en los datos de entrenamiento de cada modelo base.

2. **Selección aleatoria de características:** en cada nodo de decisión, únicamente se evalúa un subconjunto aleatorio de características candidatas para determinar la mejor división, lo que reduce la correlación entre los árboles.

La combinación de estos mecanismos permite disminuir la varianza del modelo sin incrementar de forma significativa el sesgo, lo que se traduce en mejoras consistentes en el desempeño de generalización frente a árboles de decisión individuales [111].

En el contexto del análisis de datos clínicos y sociodemográficos, Random Forest presenta múltiples ventajas relevantes:

- Alta robustez frente a valores atípicos y al ruido inherente a los datos reales.
- Capacidad para modelar relaciones no lineales y complejas entre variables clínicas y diagnósticas.
- Resistencia natural al sobreajuste gracias al esquema de *ensemble*.
- Posibilidad de estimar la importancia relativa de las características, ya sea mediante el descenso medio de impureza o a través de técnicas de *permutation importance*, lo que favorece la interpretabilidad del modelo.

Los hiperparámetros principales del modelo incluyen el número de árboles que componen el bosque (`n_estimators`), la profundidad máxima permitida para cada árbol (`max_depth`), el número mínimo de muestras requeridas para dividir un nodo interno (`min_samples_split`) y el número mínimo de muestras necesarias en un nodo hoja (`min_samples_leaf`). La selección óptima de estos hiperparámetros se llevó a cabo mediante un proceso de búsqueda en grilla (*Grid Search*) con validación cruzada estratificada, tal como se describe en la sección de entrenamiento y evaluación.

Adicionalmente, con el fin de mitigar el efecto del desbalanceo de clases presente en los datos sociodemográficos, se implementó una variante del modelo Random Forest combinada con la técnica SMOTE para el balanceo del conjunto de entrenamiento, siguiendo la misma estrategia metodológica aplicada en el modelo KNN con SMOTE. Esta aproximación permitió evaluar el impacto del balanceo de clases sobre el desempeño del clasificador sin comprometer la integridad del conjunto de prueba.

6.4. Entrenamiento de los modelos

El proceso de entrenamiento de los modelos se estructuró siguiendo las mejores prácticas documentadas en la literatura de aprendizaje automático y aprendizaje profundo [72, 69], garantizando la reproducibilidad, una monitorización adecuada del desempeño y la prevención de sobreajuste mediante estrategias de regularización y validación apropiadas.

6.4.1. Entrenamiento de los modelos de Redes Neuronales Convolucionales (CNN)

Los modelos de CNN se entrenaron de manera supervisada, ajustando iterativamente los parámetros mediante la retropropagación del error (*backpropagation*) y la optimización basada en el gradiente. Durante el proceso de entrenamiento, se empleó un esquema de mini-lotes, lo que favorece una estimación estable del gradiente y un uso eficiente de los recursos computacionales disponibles.

El desempeño del modelo se evaluó de forma periódica sobre un conjunto de validación, lo que permitió monitorear la evolución de la función de pérdida y detectar posibles indicios de sobreajuste. Los detalles específicos relacionados con la configuración de hiperparámetros, funciones de pérdida y estrategias de regularización se presentan en las subsecciones correspondientes a cada arquitectura.

6.4.1.1. Entrenamiento del modelo ResNet3D

El modelo ResNet3D se entrenó utilizando el algoritmo de optimización Adam [112], el cual combina *momentum* con tasas de aprendizaje adaptativas específicas por parámetro, favoreciendo una convergencia estable en arquitecturas profundas. Esta elección resulta especialmente adecuada para redes residuales, donde la propagación eficiente del gradiente es crítica.

La configuración de entrenamiento se definió considerando las limitaciones computacionales del entorno y la complejidad del modelo, priorizando la estabilidad y la prevención del sobreajuste. El Cuadro 6.15 presenta el conjunto de hiperparámetros empleados durante el entrenamiento del modelo.

Cuadro 6.15: Hiperparámetros de entrenamiento del modelo ResNet3D

Hiperparámetro	**Valor**
Tasa de aprendizaje inicial	1×10^{-4}
Tamaño de lote (*batch size*)	4
Número de épocas máximo	50
Optimizador	Adam ($\beta_1 = 0{,}9$, $\beta_2 = 0{,}999$)
Función de pérdida	Entropía cruzada categórica
Regularización	Dropout (0.5, solo capa final)
Inicialización de pesos	He normal

Configuración de generadores de datos

Debido a las restricciones de memoria impuestas por el tamaño volumétrico de las imágenes PET tridimensionales ($91 \times 109 \times 91$ voxeles), se implementó un sistema de generadores de datos para la

carga y el preprocesamiento dinámico de los lotes durante el entrenamiento. Esta estrategia evita la carga completa del dataset en memoria y permite un uso eficiente de los recursos computacionales disponibles.

Los generadores se encargan de leer los volúmenes desde el disco, aplicar las transformaciones necesarias y suministrar los datos en lotes de tamaño reducido al modelo. El Cuadro 6.16 resume la configuración empleada para los conjuntos de entrenamiento, validación y prueba.

Cuadro 6.16: Configuración de generadores de datos para ResNet3D

Conjunto	**Lotes por época**	**Muestras**	**Mezcla aleatoria**
Entrenamiento	1,123	4,535	Sí
Validación	251	1,018	No
Prueba	30	120	No

Arquitectura y complejidad del modelo

La arquitectura ResNet3D implementada presenta una complejidad moderada, adecuada para el procesamiento de volúmenes PET tridimensionales sin incurrir en un número excesivo de parámetros que comprometa la capacidad de generalización. El modelo cuenta con un total de 3,600,899 parámetros, de los cuales 3,598,019 son entrenables y 2,880 corresponden a parámetros no entrenables, asociados principalmente a las estadísticas de las capas de normalización por lotes.

Considerando una representación en precisión simple de 32 bits, el tamaño total del modelo es de aproximadamente 13.74 MB, lo que facilita su almacenamiento y reutilización en distintos entornos computacionales sin requerimientos de hardware especializado.

El Cuadro 6.17 detalla la distribución de los parámetros entrenables y no entrenables en cada uno de los componentes de la arquitectura, evidenciando el incremento progresivo de la capacidad representacional a medida que aumenta la profundidad de la red.

Cuadro 6.17: Distribución de parámetros en ResNet3D por componente

Componente	**Parámetros Entrenables**	**Parámetros No Entrenables**
Stem convolucional	15,776	64
Bloque residual 1 (32 filtros)	37,120	128
Bloque residual 2 (64 filtros)	148,096	256
Bloque residual 3 (128 filtros)	590,336	512
Bloque residual 4 (256 filtros)	2,359,296	1,024
Capa densa final	447,395	896
Total	**3,598,019**	**2,880**

Estrategias de regularización y monitorización

Con el objetivo de mitigar el sobreajuste y asegurar una convergencia estable durante el entrenamiento del modelo ResNet3D, se implementaron diversas estrategias de regularización y monitorización del desempeño. Estas técnicas permiten adaptar dinámicamente el proceso de optimización y seleccionar el estado del modelo con mejor capacidad de generalización.

1. **ReduceLROnPlateau:** Ajuste adaptativo de la tasa de aprendizaje, reduciéndola en un factor de 0.5 cuando la pérdida de validación no presenta mejoras durante 5 épocas consecutivas, con un umbral mínimo establecido en 1×10^{-7}. Esta estrategia facilita la convergencia fina del modelo una vez alcanzadas regiones cercanas al óptimo.

2. **ModelCheckpoint:** Almacenamiento automático del modelo con el mejor desempeño en el conjunto de validación, utilizando el *accuracy* como métrica de referencia. Este mecanismo garantiza la preservación del estado óptimo del modelo, independientemente del comportamiento en las épocas finales de entrenamiento.

3. **EarlyStopping:** Detención anticipada del entrenamiento cuando la pérdida de validación no mejora durante 10 épocas consecutivas, con restauración automática de los pesos correspondientes a la mejor época. Esta técnica previene entrenamientos innecesariamente prolongados y reduce el riesgo de sobreajuste.

6.4.1.2. Entrenamiento del modelo de Transfer Learning ResNet10-3D

El entrenamiento del modelo basado en *transfer learning* se estructuró en dos fases claramente diferenciadas, cada una con configuraciones de hiperparámetros específicas, orientadas a maximizar el aprovechamiento de las representaciones pre-entrenadas y a garantizar una adaptación progresiva al dominio de los datos PET del estudio.

Fase 1: Entrenamiento de capas superiores

En la fase inicial, correspondiente a la primera mitad del entrenamiento, es decir, de 25 épocas, se procedió al congelamiento completo del *backbone* ResNet-10 tridimensional pre-entrenado, de modo que únicamente se optimizaron los parámetros de las capas superiores diseñadas específicamente para la tarea de clasificación multiclase. Esta estrategia permite estabilizar el aprendizaje de las capas finales antes de realizar ajustes sobre las representaciones profundas del modelo.

El Cuadro 6.18 resume los hiperparámetros utilizados durante esta etapa del entrenamiento.

Cuadro 6.18: Hiperparámetros de la Fase 1 del Transfer Learning

Hiperparámetro	**Valor**
Tasa de aprendizaje	1×10^{-4}
Tamaño de lote	4
Número de épocas	25
Parámetros entrenables	282,560
Parámetros congelados	14,237,571
Optimizador	Adam

Durante esta fase, únicamente el 1.9 % de los parámetros totales del modelo fue actualizado, mientras que el 98.1 % restante, correspondiente al *backbone* pre-entrenado, permaneció fijo. Este esquema de entrenamiento favorece la preservación de las representaciones de bajo y medio nivel

aprendidas previamente a partir del conjunto de datos MedicalNet, reduciendo el riesgo de degradación temprana del conocimiento transferido.

Fase 2: Fine-tuning completo

Una vez alcanzada la convergencia de las capas superiores, se procedió al descongelamiento completo del modelo, habilitando la actualización de la totalidad de los parámetros mediante un proceso de *fine-tuning*. Esta segunda fase permite adaptar de manera controlada las representaciones profundas previamente aprendidas al dominio específico de los datos PET empleados en el estudio.

El Cuadro 6.19 resume los hiperparámetros utilizados durante esta etapa del entrenamiento.

Cuadro 6.19: Hiperparámetros de la Fase 2 del Transfer Learning (*Fine-tuning*)

Hiperparámetro	**Valor**
Tasa de aprendizaje	1×10^{-5}
Tamaño de lote	4
Número de épocas	25 (o hasta convergencia)
Parámetros entrenables	14,520,131
Parámetros congelados	0
Optimizador	Adam
Paciencia (*early stopping*)	10 épocas

La reducción de la tasa de aprendizaje en un orden de magnitud, de 1×10^{-4} a 1×10^{-5}, constituye una práctica ampliamente aceptada en procesos de *fine-tuning*, ya que permite realizar ajustes graduales sobre los pesos pre-entrenados sin comprometer de manera significativa el conocimiento previamente adquirido [102]. Esta estrategia contribuye a mejorar la capacidad de generalización del modelo y a estabilizar el proceso de optimización.

Complejidad computacional y recursos

Tras el descongelamiento completo del modelo, la arquitectura de *transfer learning* presenta un total de 14,520,131 parámetros entrenables, lo que equivale a un tamaño aproximado de 55.39 MB al emplear precisión simple de 32 bits. Esta complejidad refleja el peso computacional del *backbone* pre-entrenado, que concentra la mayor parte de los parámetros del modelo.

El Cuadro 6.20 detalla la distribución de los parámetros entre el *backbone* ResNet-10 tridimensional y las capas superiores añadidas para la tarea de clasificación específica. Como se observa, el *backbone* concentra el 98.1 % de los parámetros totales, mientras que las capas densas adicionales representan una fracción marginal del tamaño total del modelo.

Cuadro 6.20: Distribución de parámetros en el modelo Transfer Learning ResNet10-3D

Componente	Parámetros	Porcentaje	Tamaño (MB)
Backbone ResNet-10	14,237,571	98.1 %	54.31
Capa Dense 1 (512→256)	131,328	0.9 %	0.50
Capa Dense 2 (256→128)	32,896	0.2 %	0.13
Capa Dense 3 (128→3)	387	<0.1 %	<0.01
Dropout (no paramétrico)	0	0.0 %	0.00
Total	**14,520,131**	**100.0 %**	**55.39**

Esta distribución pone de manifiesto una característica fundamental del enfoque de *transfer learning*: la reutilización de representaciones profundas previamente aprendidas permite incorporar conocimiento relevante con un costo computacional relativamente bajo en las capas específicas de la tarea, concentrando la complejidad en componentes ya entrenados sobre grandes volúmenes de datos.

Monitorización y criterios de detención

Durante ambas fases del entrenamiento del modelo de *transfer learning*, se emplearon los mismos mecanismos de monitorización y control descritos previamente para el modelo ResNet3D. La utilización de criterios consistentes permite una comparación metodológicamente válida entre ambos enfoques.

En particular, se implementaron las siguientes estrategias:

- Reducción adaptativa de la tasa de aprendizaje cuando la pérdida de validación no presenta mejoras durante un número predefinido de épocas.
- Guardado automático del modelo con mejor desempeño en el conjunto de validación, utilizando el *accuracy* como métrica de referencia.
- Detención anticipada del entrenamiento tras 10 épocas consecutivas sin mejora en la pérdida de validación, con restauración de los mejores pesos.

La aplicación de este esquema de monitorización en dos fases permite maximizar el aprovechamiento de las representaciones pre-entrenadas y, simultáneamente, ajustar de forma controlada el modelo a la tarea específica de clasificación de la enfermedad de Alzheimer a partir de imágenes PET tridimensionales.

6.4.1.3. Entrenamiento del modelo VoxCNN3D

El modelo VoxCNN3D fue entrenado priorizando la eficiencia computacional y la estabilidad del proceso de optimización, sin comprometer su capacidad para extraer características discriminativas a partir de datos volumétricos. Esta arquitectura fue concebida como una alternativa ligera frente a modelos más profundos.

El Cuadro 6.21 resume los principales hiperparámetros utilizados durante el entrenamiento:

Cuadro 6.21: Hiperparámetros de entrenamiento del modelo VoxCNN3D

Hiperparámetro	**Valor**
Tasa de aprendizaje inicial	3×10^{-4}
Tamaño de lote (*batch size*)	8
Número de épocas máximo	50
Optimizador	Adam ($\beta_1 = 0{,}9$, $\beta_2 = 0{,}999$)
Función de pérdida	Entropía cruzada categórica
Regularización	Dropout progresivo (0.15-0.30)
Inicialización de pesos	He normal

En primer lugar, se empleó el optimizador Adam con una tasa de aprendizaje inicial moderada, junto con la función de pérdida de entropía cruzada categórica, adecuada para tareas de clasificación multiclase. Adicionalmente, se incorporó regularización mediante *dropout* progresivo para mitigar el sobreajuste y una inicialización de pesos tipo He normal, apropiada para arquitecturas basadas en funciones de activación ReLU.

El número máximo de épocas se estableció en 50, y el proceso de entrenamiento está sujeto a los criterios de monitoreo y detención anticipada descritos posteriormente.

Configuración de generadores y eficiencia computacional

Una de las principales ventajas del modelo VoxCNN3D radica en su menor complejidad arquitectónica, lo que permite procesar lotes de mayor tamaño (*batch size* = 8) en comparación con ResNet3D (*batch size* = 4), sin exceder las limitaciones de memoria de la CPU. Esta característica se traduce en una reducción significativa del número de lotes por época y, por ende, en una mejora de la eficiencia computacional global del entrenamiento.

El Cuadro 6.22 presenta la configuración de los generadores de datos para los conjuntos de entrenamiento, validación y prueba. En el conjunto de entrenamiento, se aplicó una mezcla aleatoria de las muestras (*shuffling*) con el fin de favorecer la generalización del modelo, mientras que en los conjuntos de validación y prueba se mantuvo un orden fijo para garantizar la reproducibilidad de los resultados.

Cuadro 6.22: Configuración de generadores de datos para VoxCNN3D

Conjunto	**Lotes/época**	**Muestras**	**Tiempo/época**	**Mezcla**
Entrenamiento	562	4,535	∼122 min	Sí
Validación	126	1,018	∼18 min	No
Prueba	15	120	∼2 min	No

El incremento en el tamaño del lote permitió reducir aproximadamente a la mitad el número de iteraciones por época en comparación con ResNet3D, lo cual se reflejó en una disminución del tiempo total de entrenamiento por época.

Arquitectura y complejidad del modelo

El modelo VoxCNN3D fue diseñado con un enfoque en la eficiencia computacional, priorizando una arquitectura compacta capaz de capturar patrones relevantes en datos volumétricos sin incurrir

en un elevado costo de memoria. La arquitectura completa comprende un total de 1,277,123 parámetros, de los cuales 1,275,987 son entrenables y 1,136 corresponden a parámetros no entrenables asociados a las estadísticas de normalización por lotes.

El tamaño total del modelo es de aproximadamente 4.87 MB en precisión de 32 bits, lo que representa una reducción sustancial frente a arquitecturas más profundas como ResNet3D de 13.74 MB y el modelo de Transfer Learning ResNet10-3D de 55.39 MB.

El Cuadro 6.23 presenta la distribución de parámetros por componente arquitectónico, evidenciando que la mayor proporción de parámetros entrenables se concentra en el clasificador totalmente conectado, mientras que los bloques convolucionales mantienen una complejidad moderada gracias al uso progresivo de filtros y a la normalización por lotes.

Cuadro 6.23: Distribución de parámetros en VoxCNN3D por componente

Componente	**Parámetros Entrenables**	**Parámetros No Entrenables**
Stem convolucional (24 filtros)	25,944	48
Bloque 1 (48 filtros)	52,416	192
Bloque 2 (96 filtros)	208,512	384
Bloque 3 (128 filtros)	369,280	512
Clasificador Dense (192 unidades)	619,835	0
Total	**1,275,987**	**1,136**

Estrategia de regularización anti-colapso

La estrategia de regularización adoptada para VoxCNN3D fue diseñada específicamente para mitigar el colapso del modelo hacia la predicción dominante de una única clase, un comportamiento identificado en versiones preliminares bajo esquemas de regularización excesivamente restrictivos. En este contexto, se priorizó un equilibrio entre la capacidad de representación y el control del sobreajuste, lo que dio lugar a las siguientes decisiones metodológicas:

1. **Tasa de aprendizaje elevada:** Se empleó una tasa de aprendizaje inicial de 3×10^{-4}, superior a la utilizada en ResNet3D, con el objetivo de favorecer una exploración más amplia del espacio de parámetros y reducir la probabilidad de convergencia prematura hacia mínimos locales subóptimos.

2. **Dropout progresivo moderado:** Se implementó un esquema de dropout creciente ($0.15 \rightarrow 0.20 \rightarrow 0.25 \rightarrow 0.30$) a lo largo de la arquitectura, evitando valores uniformemente altos que pueden limitar la capacidad de aprendizaje, especialmente en las capas iniciales.

3. **Ausencia de regularización L2:** No se aplicó penalización por norma L2 (*weight decay*) con el fin de evitar una restricción excesiva sobre los pesos del modelo, lo cual podría favorecer el colapso en escenarios de datos desbalanceados.

4. **Stem convolucional con reducción espacial temprana:** Se utilizó un *stride* de 2 en la capa inicial para realizar una reducción espacial temprana del volumen de entrada, disminuyendo la complejidad computacional y contribuyendo a una convergencia más estable.

Callbacks y monitorización

Para garantizar la estabilidad del proceso de entrenamiento y una convergencia adecuada, se emplearon los mismos mecanismos de monitorización utilizados en los modelos previamente descritos, ajustando únicamente los parámetros de paciencia en función de la dinámica de aprendizaje de VoxCNN3D. En particular, se implementaron los siguientes *callbacks*:

- **ReduceLROnPlateau:** Reducción automática de la tasa de aprendizaje por un factor de 0.5 cuando la pérdida de validación no presenta mejoras durante 5 épocas consecutivas.
- **EarlyStopping:** Detención anticipada del entrenamiento tras 10 épocas sin mejora en la pérdida de validación, con restauración automática de los pesos correspondientes al mejor desempeño.
- **ModelCheckpoint:** Almacenamiento automático del modelo con mayor *accuracy* en el conjunto de validación, asegurando la conservación del mejor estado del modelo independientemente de la época final.

6.4.2. Entrenamiento de los Modelos de Machine Learning (ML)

El entrenamiento de los modelos de aprendizaje automático clásico se llevó a cabo siguiendo un enfoque sistemático de optimización y evaluación, orientado a garantizar un desempeño robusto y comparable entre algoritmos. Para ello, se emplearon esquemas de búsqueda de hiperparámetros combinados con validación cruzada estratificada, lo que permitió seleccionar configuraciones adecuadas y obtener estimaciones confiables de la capacidad de generalización de los modelos. Los detalles específicos del proceso de entrenamiento y los hiperparámetros finales seleccionados se describen en los apartados correspondientes a cada modelo.

6.4.2.1. Entrenamiento del modelo K-Nearest Neighbors (KNN)

El entrenamiento del modelo *K-Nearest Neighbors* (KNN) se centró en la selección del valor óptimo del hiperparámetro k, correspondiente al número de vecinos considerados en el proceso de clasificación. Para ello, se evaluó un rango predefinido de valores de k mediante validación cruzada estratificada, garantizando un balance adecuado entre sesgo y varianza. La configuración final del modelo se seleccionó con base en el desempeño promedio obtenido durante la validación, asegurando una estimación robusta de su capacidad de generalización.

Estrategia de Búsqueda de hiperparámetros

La optimización del modelo KNN se realizó mediante una búsqueda exhaustiva del hiperparámetro k, evaluando valores en el rango $[1, 30]$. Para cada configuración, el desempeño se estimó utilizando validación cruzada estratificada, seleccionando finalmente el valor de k que maximizó el *accuracy* promedio en el conjunto de validación. El Cuadro 6.24 resume el espacio de búsqueda considerado y la configuración final adoptada.

Cuadro 6.24: Espacio de búsqueda de hiperparámetros para KNN

Hiperparámetro	Rango/Valores	Configuración Final
Número de vecinos (k)	1 a 30	Determinado empíricamente
Métrica de distancia	Euclidiana	Euclidiana
Esquema de ponderación	Uniforme	Uniforme

Validación cruzada

Para cada valor de k considerado, se aplicó un esquema de validación cruzada estratificada de 5 pliegues sobre el conjunto de entrenamiento, calculando el *accuracy* promedio y su correspondiente desviación estándar. Este enfoque permite obtener una estimación más robusta y estable del desempeño del modelo, en comparación con una única partición entre entrenamiento y validación, lo cual resulta especialmente relevante dada la naturaleza y el tamaño moderado del conjunto de datos.

Asimismo, la validación cruzada garantiza que cada muestra del conjunto de entrenamiento se utilice una vez como datos de validación y en los cuatro pliegues restantes como datos de entrenamiento, optimizando el uso de la información disponible y reduciendo la varianza asociada a la estimación del desempeño.

6.4.2.2. Entrenamiento del modelo K-Nearest Neighbors (KNN) con SMOTE

El entrenamiento del modelo KNN con SMOTE incorporó una etapa adicional de balanceo de clases con el objetivo de mitigar el desbalance presente en el conjunto de datos. Dicho proceso se implementó mediante un *pipeline* de preprocesamiento, garantizando que todas las transformaciones se aplicaran de manera consistente y exclusivamente sobre los datos de entrenamiento, evitando así la introducción de sesgos por fuga de información.

El Cuadro 6.25 resume la secuencia de transformaciones aplicadas durante el entrenamiento. En una primera etapa, se realizó la imputación de valores faltantes utilizando la mediana, seguida de la estandarización de las variables para asegurar una escala homogénea, condición especialmente relevante para el funcionamiento adecuado del algoritmo KNN. Posteriormente, se aplicó el método SMOTE para la generación sintética de muestras de las clases minoritarias, utilizando $k_{neighbors} = 5$. Finalmente, se entrenó el clasificador KNN empleando el valor óptimo de k determinado en la etapa de validación cruzada.

Cuadro 6.25: Pipeline de preprocesamiento para KNN con SMOTE

Etapa	Transformación	Parámetros
1	Imputación + Estandarización	Mediana, $\mu = 0$, $\sigma = 1$
2	SMOTE	$k_{neighbors} = 5$
3	Clasificador KNN	k optimizado

Optimización de hiperparámetros con GridSearchCV

Con el fin de identificar la configuración óptima del modelo KNN con SMOTE, se implementó una búsqueda exhaustiva de hiperparámetros mediante `GridSearchCV`, utilizando validación cruzada

estratificada de 5 pliegues. En comparación con el modelo KNN base, el espacio de búsqueda se amplió para explorar distintas combinaciones que afectan tanto la definición del vecindario como el cálculo de distancias.

El Cuadro 6.26 presenta el espacio de búsqueda considerado, el cual incluyó variaciones en el número de vecinos, el esquema de ponderación, la métrica de distancia y el parámetro p de la distancia de Minkowski.

Cuadro 6.26: Espacio de búsqueda extendido para KNN con SMOTE

Hiperparámetro	**Valores Evaluados**
`n_neighbors`	[3, 5, 7, 9, 11, 15]
`weights`	['uniform', 'distance']
`metric`	['euclidean', 'manhattan', 'minkowski']
`p` (parámetro Minkowski)	[1, 2]
Total de combinaciones	**72**

En total, se evaluaron 72 combinaciones de hiperparámetros, cada una validada mediante 5 pliegues, lo que resultó en 360 ajustes de modelo. La selección final de los hiperparámetros se realizó con base en el *accuracy* promedio obtenido durante la validación cruzada, garantizando una estimación robusta del desempeño del modelo bajo distintas configuraciones.

6.4.2.3. Entrenamiento del modelo Naive Bayes

El entrenamiento del clasificador Naive Bayes en su variante Gaussiana no requiere un proceso explícito de optimización de hiperparámetros, dado que el modelo estima de forma directa los parámetros de las distribuciones de probabilidad asociadas a cada característica. En particular, las medias y varianzas de las distribuciones gaussianas se calculan mediante máxima verosimilitud a partir del conjunto de entrenamiento.

Estimación de parámetros

Durante el entrenamiento del modelo Naive Bayes Gaussiano, se estimaron los parámetros estadísticos necesarios para caracterizar la distribución de cada característica continua condicionada a la clase. Para cada característica x_i y cada clase y_j, el modelo calcula la media y la varianza utilizando estimadores de máxima verosimilitud.

Adicionalmente, las probabilidades a priori de cada clase $P(y_j)$ se estimaron a partir de las frecuencias empíricas observadas. El Cuadro 6.27 resume el conjunto de parámetros estimados por el modelo, evidenciando su baja complejidad y eficiencia computacional.

Cuadro 6.27: Parámetros estimados por el modelo Naive Bayes

Componente	**Cantidad**	**Método de Estimación**
Medias (μ_{ij})	$11 \times 3 = 33$	Máxima verosimilitud
Varianzas (σ^2_{ij})	$11 \times 3 = 33$	Máxima verosimilitud
Probabilidades a priori $P(y_j)$	3	Frecuencias empíricas
Total de parámetros	**69**	-

Cabe resaltar que, aunque no se optimizaron los hiperparámetros, se realizó una validación cruzada estratificada de 5 pliegues para obtener una estimación robusta del desempeño del modelo.

6.4.2.4. Entrenamiento del modelo Random Forest

El entrenamiento del modelo Random Forest se llevó a cabo mediante un proceso sistemático de optimización de hiperparámetros, el cual es inherentemente más complejo en comparación con modelos como KNN o Naive Bayes. Esta complejidad se debe a la mayor cantidad de parámetros configurables que controlan tanto la estructura individual de los árboles de decisión como el comportamiento global del *ensemble*, influyendo directamente en el equilibrio entre sesgo, varianza y capacidad de generalización.

Espacio de búsqueda de hiperparámetros

Con el objetivo de identificar la configuración óptima del modelo, se definió un espacio de búsqueda multidimensional que abarca los hiperparámetros con mayor influencia sobre la capacidad predictiva y la complejidad del Random Forest. Estos hiperparámetros controlan tanto la diversidad del conjunto de árboles como la profundidad y granularidad de las particiones individuales, afectando directamente el balance entre sesgo y varianza.

El Cuadro 6.28 resume el espacio completo de búsqueda considerado durante el proceso de optimización.

Cuadro 6.28: Espacio de búsqueda de hiperparámetros para Random Forest

Hiperparámetro	Valores Evaluados	Descripción
`n_estimators`	[50, 100, 200, 300]	Número de árboles
`max_depth`	[None, 10, 20, 30]	Profundidad máxima
`min_samples_split`	[2, 5, 10]	Min. muestras para dividir
`min_samples_leaf`	[1, 2, 4]	Min. muestras en hoja
`max_features`	['sqrt', 'log2']	Features por división
`bootstrap`	[True, False]	Uso de bootstrap
Total de combinaciones:	$4 \times 4 \times 3 \times 3 \times 2 \times 2 = \mathbf{576}$	

Optimización con GridSearchCV

La optimización de hiperparámetros se llevó a cabo mediante `GridSearchCV`, empleando validación cruzada estratificada de 5 pliegues. Dado el espacio de búsqueda definido, este procedimiento implicó un total de $576 \times 5 = 2{,}880$ entrenamientos independientes del modelo, garantizando una evaluación exhaustiva de cada configuración candidata.

Con el fin de mitigar la carga computacional asociada, se habilitó la paralelización completa del proceso (`n_jobs=-1`), permitiendo la distribución simultánea de las evaluaciones entre todos los núcleos de CPU disponibles.

La selección del modelo óptimo se basó en la maximización del *accuracy* medio obtenido durante la validación cruzada. La mejor puntuación alcanzada fue registrada en el atributo `best_score_`, mientras que la combinación de hiperparámetros correspondiente se almacenó en `best_params_`.

Random Forest con SMOTE

De forma análoga a los modelos basados en KNN, se implementó una variante del clasificador Random Forest que incorpora una estrategia de balanceo de clases mediante SMOTE, con el objetivo de mitigar el sesgo inducido por la distribución desbalanceada de las clases diagnósticas.

El proceso se estructuró mediante un *pipeline* integrado que garantiza la aplicación consistente de las etapas de preprocesamiento, balanceo y clasificación. El Cuadro 6.29 resume la secuencia de transformaciones aplicadas a los datos:

Cuadro 6.29: Pipeline de preprocesamiento para Random Forest con SMOTE

Etapa	Transformación	Parámetros
1	Imputación + Estandarización	Mediana, $\mu = 0$, $\sigma = 1$
2	SMOTE	$k_{neighbors} = 5$
3	Random Forest	Hiperparámetros optimizados

La optimización de hiperparámetros se realizó sobre el mismo espacio de búsqueda definido para el modelo de Random Forest base. No obstante, en este caso, el procedimiento de SMOTE se aplicó de manera exclusiva sobre los datos de entrenamiento dentro de cada pliegue de la validación cruzada, evitando así la fuga de información y garantizando una evaluación realista del desempeño de generalización del modelo.

6.5. Estrategias de evaluación de los modelos implementados

Todos los modelos implementados, tanto de aprendizaje profundo como de aprendizaje automático tradicional, se evaluaron utilizando un conjunto consistente de métricas de rendimiento que permiten la comparación sistemática entre arquitecturas. La selección de métricas responde a la naturaleza multiclase del problema de clasificación y a la importancia clínica de detectar correctamente todas las categorías diagnósticas, particularmente las clases minoritarias.

Para los modelos de aprendizaje profundo, se monitorizaron adicionalmente las curvas deL accuracy durante el entrenamiento y la validación, lo que permitió diagnosticar problemas de sobreajuste, subajuste o convergencia prematura.

6.5.1. Protocolo de validación

Todos los modelos siguieron un protocolo de evaluación de tres etapas:

1. **Validación durante el entrenamiento:** Evaluación periódica sobre el conjunto de validación para monitorear la convergencia y activar mecanismos de early stopping o reducción de learning rate.

2. **Validación cruzada (solo modelos de ML):** Estimación de la variabilidad del rendimiento mediante una validación cruzada estratificada de 5 pliegues sobre el conjunto de entrenamiento.

3. **Evaluación final en test:** Evaluación única sobre el conjunto de prueba reservado, nunca utilizado durante el entrenamiento ni la selección de hiperparámetros, proporcionando una estimación imparcial del rendimiento de generalización.

Este protocolo riguroso garantiza que las métricas reportadas reflejan genuinamente la capacidad de los modelos para generalizar a datos no vistos, un criterio fundamental para su aplicabilidad clínica potencial.

6.6. Consideraciones sobre fuga de información

Un aspecto crítico en el diseño del script de preprocesamiento y entrenamiento fue la prevención de la fuga de información (*data leakage*) [94]. Este fenómeno ocurre cuando información proveniente del conjunto de prueba influye directa o indirectamente en el proceso de entrenamiento, dando lugar a estimaciones artificialmente optimistas y no representativas del desempeño de generalización del modelo.

Con el fin de mitigar este riesgo y garantizar una evaluación metodológicamente rigurosa, se adoptaron las siguientes estrategias:

1. **Ajuste de transformadores exclusivamente en el conjunto de entrenamiento:** Los parámetros de todos los transformadores empleados en el preprocesamiento, incluyendo la mediana para imputación y la media y desviación estándar para estandarización, se estimaron únicamente a partir del conjunto de entrenamiento mediante el método `fit()`. Posteriormente, dichos parámetros se aplicaron al conjunto de validación y prueba utilizando `transform()`, evitando cualquier recalibración basada en datos no vistos.

2. **Aplicación de SMOTE limitada al conjunto de entrenamiento:** En los modelos que incorporaron técnicas de balanceo de clases mediante SMOTE, la generación de muestras sintéticas se realizó exclusivamente sobre los datos de entrenamiento. El conjunto de prueba conservó su distribución original de clases, permitiendo evaluar el desempeño del modelo bajo condiciones realistas de desbalanceo propias del contexto clínico.

3. **Integración adecuada de SMOTE en validación cruzada:** Para los escenarios que combinaron validación cruzada con balanceo de clases, se utilizó `ImbPipeline` de la librería *imbalanced-learn*, en lugar del `Pipeline` estándar. Esta implementación garantiza que SMOTE se aplique de forma independiente en cada pliegue de entrenamiento durante la validación cruzada, sin contaminar los pliegues de validación.

La adopción de estas prácticas metodológicas asegura que las métricas de desempeño obtenidas reflejen de manera fiel la capacidad de generalización de los modelos ante datos no vistos, proporcionando estimaciones válidas, reproducibles y confiables.

6.6.1. Balanceo de clases mediante SMOTE

Con el fin de mitigar el desbalanceo significativo entre las clases diagnósticas presentes en los datos tabulares, se incorporó una estrategia de sobremuestreo sintético en los modelos K-Nearest Neighbors con SMOTE y Random Forest con SMOTE. Para este propósito, se empleó la técnica SMOTE (*Synthetic Minority Over-sampling Technique*) [106], ampliamente utilizada en problemas de clasificación médica con clases minoritarias.

SMOTE genera nuevas instancias sintéticas de las clases minoritarias mediante interpolación lineal entre muestras existentes y sus k vecinos más cercanos en el espacio de características, incrementando la representación de dichas clases sin recurrir a la simple duplicación de ejemplos.

6.6.1.1. Configuración de SMOTE

La implementación de SMOTE se realizó utilizando una configuración estándar y reproducible, definida por los siguientes parámetros:

- `k_neighbors = 5`**:** número de vecinos más cercanos considerados para la generación de muestras sintéticas.
- `random_state = 42`**:** semilla fija para garantizar la reproducibilidad de los experimentos.
- **Estrategia de balanceo:** igualación del número de muestras de todas las clases al tamaño de la clase mayoritaria.

La aplicación de SMOTE sobre el conjunto de entrenamiento original, compuesto por 3,693 muestras.

6.7. Modelo Híbrido: Integración de Transfer Learning con ResNet10-3D y Random Forest con estadísticas clínicas poblacionales y estudios PET

En este trabajo se propone un modelo híbrido que integra información proveniente de dos fuentes distintas: estudios PET cerebrales y datos clínicos y sociodemográficos. No obstante, es importante aclarar que dicha integración no se realiza a nivel individual por paciente, debido a que ambas fuentes de información presentan diferencias en el número de registros disponibles y en la distribución de las categorías diagnósticas.

Los estudios PET corresponden a sujetos individuales. Por el contrario, los datos clínicos disponibles no permiten una correspondencia directa uno a uno con los estudios PET, ya que el número de registros clínicos es menor y no existe una coincidencia exacta entre sujetos ni entre categorías diagnósticas.

Por esta razón, el modelo propuesto no realiza una fusión sujeto–sujeto entre estudios PET y datos clínicos individuales, ya que dicha estrategia sería metodológicamente incorrecta y podría introducir asociaciones erróneas entre pacientes.

6.7.1. Uso de estadísticas clínicas poblacionales

En lugar de utilizar datos clínicos individuales, se optó por emplear estadísticas clínicas poblacionales, calculadas a partir del conjunto completo de datos clínicos disponibles. Estas estadísticas describen el comportamiento general de distintas variables clínicas para cada uno de los grupos diagnósticos considerados.

Para cada variable clínica disponible (edad, años de educación, puntaje MMSE, presencia del alelo APOE-ε4, género, peso) y para cada una de las tres categorías diagnósticas, se calculan cinco estadísticos descriptivos sobre el conjunto completo de datos clínicos:

- Media aritmética (μ), desviación estándar (σ), mediana, valor mínimo y máximo.

Este proceso genera un vector único de características estadísticas de dimensión m:

$$m = n_{\text{variables}} \times n_{\text{categorías}} \times n_{\text{estadísticos}} = n_{\text{variables}} \times 3 \times 5 \tag{6.2}$$

El vector resultante $\mathbf{v}_{\text{stats}} \in \mathbb{R}^m$ encapsula conocimiento poblacional sobre cómo se distribuyen las variables clínicas en cada grupo diagnóstico.

Este conjunto de estadísticas conforma un vector clínico global que representa información contextual sobre la enfermedad y sus manifestaciones clínicas a nivel poblacional, y no sobre un paciente en particular. Dicho vector estadístico es único y se mantiene constante para todos los estudios PET analizados.

6.7.2. Extracción de información a partir de estudios PET

Cada estudio PET es procesado de forma individual mediante un modelo profundo previamente entrenado basado en una arquitectura ResNet 3D. Este modelo no se utiliza como clasificador final, sino como un extractor de características, cuyo objetivo es resumir la información relevante contenida en los estudios.

Como resultado de este proceso, cada estudio PET es transformado en un vector numérico de características que describe los patrones metabólicos cerebrales aprendidos por el modelo durante su entrenamiento previo. Este vector constituye una representación individual del estado cerebral del sujeto correspondiente al estudio PET.

6.7.3. Proceso de concatenación de la información

La integración de ambas fuentes de información se realiza mediante un proceso de concatenación, en el cual se combinan, para cada estudio PET, dos componentes:

1. La representación individual extraída del estudio PET.
2. El vector de estadísticas clínicas poblacionales, que actúa como información contextual.

El vector de estadísticas clínicas se replica por cada estudio PET que se procesa, de manera que cada estudio se asocia con el mismo contexto clínico global. De este modo, la concatenación

no implica una correspondencia directa entre sujetos de ambas modalidades, sino una combinación entre información individual (estudio PET) y conocimiento poblacional (variables clínicas).

Este enfoque permite integrar información clínica sin requerir que ambas fuentes tengan el mismo número de registros ni una correspondencia exacta por sujeto.

6.7.4. Clasificación final

La información concatenada se utiliza como entrada para un clasificador basado en Random Forest, el cual aprende a discriminar entre las categorías diagnósticas, considerando tanto los patrones visuales extraídos de los estudios PET como el contexto clínico global proporcionado por las estadísticas poblacionales.

De esta forma, la decisión final del modelo se basa principalmente en la información individual contenida en cada estudio, mientras que la información clínica contribuye como un apoyo contextual que puede mejorar la robustez del proceso de clasificación.

6.7.5. Arquitectura del modelo híbrido

La Figura 6.3 ilustra esquemáticamente la arquitectura completa del sistema.

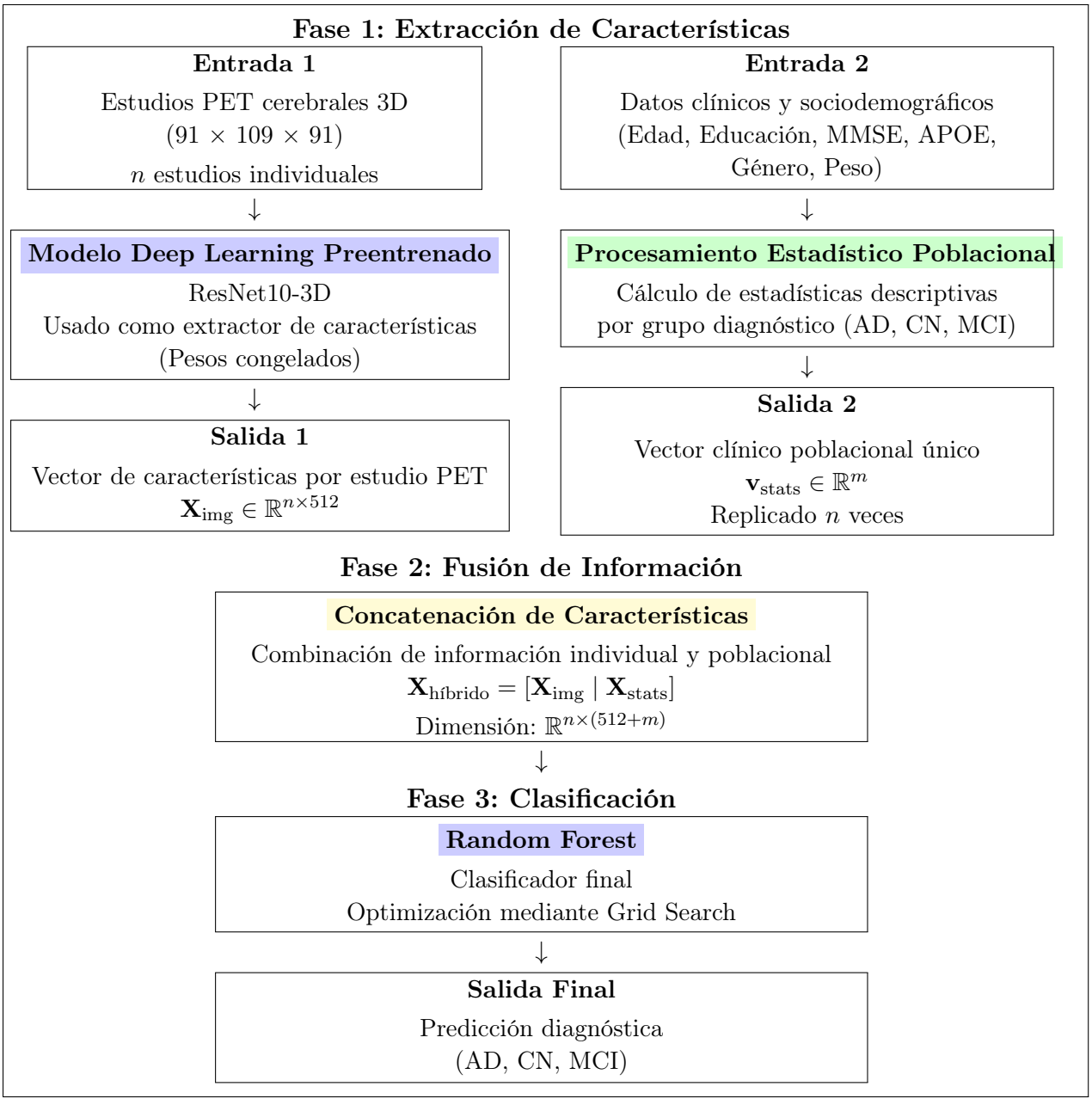


Figura 6.3: Arquitectura del modelo híbrido propuesto. El sistema combina información individual extraída de estudios PET mediante un modelo Deep Learning preentrenado usado como extractor de características, con información clínica poblacional representada mediante estadísticas descriptivas. El vector clínico global se replica para cada estudio PET, permitiendo la fusión por concatenación sin establecer correspondencia sujeto–sujeto.

6.8. Interfaz gráfica de visualización

Con el objetivo de facilitar la interacción con el sistema de clasificación diagnóstica, se desarrolló una interfaz gráfica de usuario (GUI) mediante el framework Gradio 4.x [113].

6.8.1. Arquitectura de la interfaz

La interfaz se estructuró como una aplicación web interactiva, implementada completamente en Python y ejecutable en entornos locales o mediante un túnel público. Se organizó en cinco módulos funcionales:

1. **Sistema de autenticación:** Control de acceso basado en credenciales de usuario.
2. **Módulo de preprocesamiento:** Pipeline automatizado que ejecuta las etapas descritas en la Sección 6.2.1, incluyendo la conversión de formatos, el registro espacial, la normalización de intensidades y la aplicación de la máscara cerebral.
3. **Motor de inferencia:** Componente que carga el modelo híbrido y ejecuta la predicción, integrando características de neuroimagen y datos clínicos.
4. **Generador de reportes:** Sistema de generación automática de reportes en formatos TXT y PDF.
5. **Sistema de visualización:** Módulo de renderizado interactivo mediante Plotly [114].

6.8.2. Generación de reportes clínicos

El sistema genera reportes mediante `generate_text_report()`, que construye un documento estructurado en nueve secciones:

Ítem	Sección del Reporte
1	Encabezado del sistema con advertencia sobre su carácter prototípico.
2	Información demográfica del paciente.
3	Evaluación cognitiva con interpretación de MMSE.
4	Información genética (APOE) y antropométrica.
5	Hábitos y factores de riesgo vascular.
6	Resultado del análisis diagnóstico con probabilidades.
7	Análisis automatizado de factores de riesgo.
8	Recomendaciones clínicas personalizadas.
9	Limitaciones y disclaimer legal extenso.

Cuadro 6.30: Estructura del reporte clínico generado por el sistema

6.8.3. Visualización multiplanar

El módulo de visualización implementa dos modos de renderizado basados en Plotly:

Visualización de plano único. La función `create_3d_brain_visualization()` genera vistas 2D de planos axial, coronal o sagital. Para cada plano, se calcula el índice de corte central y se utiliza el mapa de color "Hot", optimizado para estudios PET.

Visualización multiplanar simultánea. La función `create_multiplanar_visualization()` genera una figura con un subplot de 1 fila × 3 columnas, mostrando simultáneamente los tres planos ortogonales y facilitando la evaluación espacial completa del patrón metabólico.

6.8.4. Pestañas de la interfaz de usuario

El Cuadro 6.31 muestra la estructura de la interfaz mediante bloques y pestañas de Gradio, organizando el flujo en etapas lógicas:

Pestaña	**Componentes**	**Funcionalidad**
1. Sistema de autenticación	`verificar_login()`	Compara las credenciales contra un diccionario predefinido.
2. Información Personal	`gr.Textbox()`, `gr.Number()`, `gr.Radio()`	Captura de datos demográficos del paciente.
3. Evaluación Cognitiva	`gr.Slider()`, `gr.Dropdown()`	MMSE (rango 0-30) y selección de estado cognitivo previo.
4. Genética y Antropometría	`gr.Dropdown()`, `gr.Number()`	Selección de genotipo APOE, entrada de peso/altura, cálculo automático de IMC.
5. Hábitos y Comorbilidades	`gr.Checkbox()`	Selección de variables binarias (hábitos y condiciones médicas).
6. Estudio PET	`gr.File()`, `gr.Dropdown()`, `gr.Plot()`	Carga de imágenes médicas, selector de vista anatómica, previsualización.
7. Diagnóstico y Resultados	`gr.Plot()`, `gr.Button()`, `gr.File()`	Visualización de resultados, gráficos de diagnóstico, descarga de reportes.

Cuadro 6.31: Estructura de pestañas con componentes y funcionalidades de la interfaz

6.9. Descripción de las herramientas de software y hardware implementadas

El desarrollo, el entrenamiento y la evaluación de los modelos de aprendizaje automático y de las redes neuronales convolucionales se realizaron utilizando un sistema computacional con las siguientes especificaciones técnicas:

- **Procesador:** Intel Core i5-12400F.

- **Memoria RAM:** 16 GB DDR4 a 3200 MHz.
- **Almacenamiento interno:** Unidad de estado sólido (SSD) NVMe con una capacidad de 2 TB.
- **Sistema Operativo:** Windows 11 de 64 bits.
- **Tarjeta gráfica:** AMD Radeon RX 6600.
- **Disco duro externo:** Capacidad de 1 TB para el almacenamiento y respaldo de los conjuntos de datos y resultados experimentales.

6.9.1. Entorno de programación y librerías

Para el desarrollo del proyecto se empleó Python [115] como lenguaje principal. En el Cuadro 6.32 se presenta un resumen de las principales librerías utilizadas, organizadas por categoría funcional.

Cuadro 6.32: Librerías principales empleadas en el proyecto

Categoría	**Librería**
Manipulación de datos	NumPy [116] Pandas [117] Openpyxl [118]
Aprendizaje automático	Scikit-learn [119] Imbalanced-learn [120]
Visualización	Matplotlib [121] Seaborn [122]
Persistencia	Pickle [123] Joblib [124]
Deep learning y neuroimágenes	TensorFlow/Keras [125] PyTorch [126] NiBabel [127] SciPy [128] SimpleITK [129]
Interfaz gráfica	Gradio [113] Plotly [114] ReportLab [130]
Sistema y utilidades	Os [131] Json [132] Warnings [133] Datetime [134]

Capítulo 7

Resultados y Discusión

En esta sección se presentan y analizan los resultados obtenidos a partir de la evaluación del desempeño de los modelos de clasificación implementados. Por su parte, el análisis se fundamenta en dos tipos de evidencia: datos experimentales obtenidos a partir de la medición directa de métricas de rendimiento durante las fases de entrenamiento, validación y prueba; y datos empíricos derivados del análisis descriptivo y la observación de las propiedades estadísticas y visuales del conjunto de datos utilizado.

Esta distinción permite complementar la evaluación cuantitativa del desempeño de los modelos con una interpretación cualitativa de los patrones presentes en los datos, proporcionando una visión integral de los resultados y de su relevancia para el problema de clasificación abordado.

7.1. Análisis exploratorio del conjunto de datos de imágenes PET

Previo al entrenamiento de los modelos de aprendizaje profundo, se llevó a cabo un análisis exploratorio del conjunto de datos de imágenes PET con el fin de caracterizar visualmente las principales diferencias entre las categorías diagnósticas consideradas y verificar la adecuación del preprocesamiento aplicado.

7.1.1. Características visuales por categoría diagnóstica

Las Figuras 7.1, 7.2 y 7.3 presentan ejemplos representativos de cortes axiales de imágenes PET para las tres categorías diagnósticas: Enfermedad de Alzheimer (AD), Control Normal (CN) y Deterioro Cognitivo Leve (MCI), respectivamente.

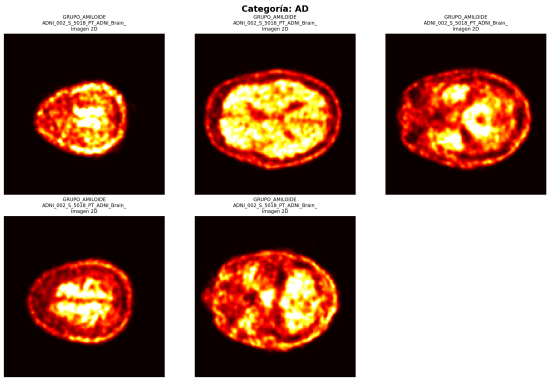


Figura 7.1: Ejemplos representativos de imágenes PET para la categoría AD.

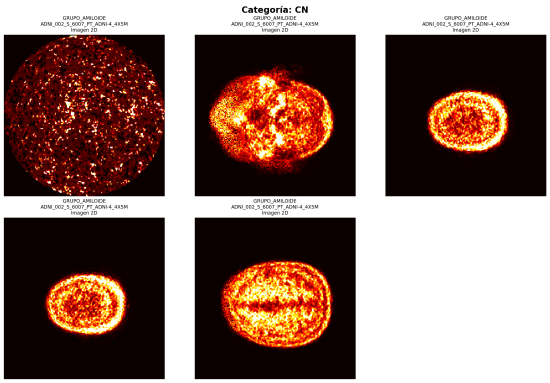


Figura 7.2: Ejemplos representativos de imágenes PET para la categoría CN.

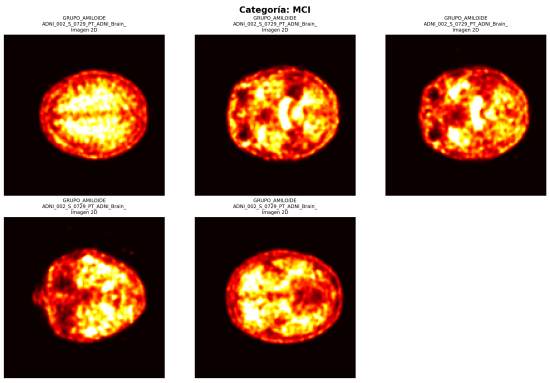


Figura 7.3: Ejemplos representativos de imágenes PET para la categoría MCI.

7.1.1.1. Observaciones cualitativas por categoría

Las imágenes correspondientes a la categoría AD (Figura 7.1) exhiben un patrón característico de captación cortical difusa y elevada del trazador amiloide, manifestándose como regiones de alta intensidad correspondientes a los tonos amarillo-blanco, distribuidas a lo largo del manto cortical. Este patrón espacial es consistente con la deposición extensa de placas de β-amiloide documentada en estudios neuropatológicos y de neuroimagen en pacientes con enfermedad de Alzheimer [36]. La elevada intensidad de señal, junto con su relativa homogeneidad cortical, constituye un marcador visual robusto de positividad amiloide y un rasgo distintivo frente a otras categorías diagnósticas.

En contraste, las imágenes de la categoría CN (Figura 7.2) presentan una captación cortical mínima o prácticamente ausente, evidenciada por la predominancia de tonalidades rojizas y naranjas asociadas a una baja retención del trazador. En algunos casos, se observa un patrón de ruido granular o punteado más pronunciado, atribuible a la naturaleza estocástica de la emisión de positrones en la modalidad PET. Este ruido cuántico, inherente a la técnica, se manifiesta en grados variables y no compromete la capacidad diagnóstica cuando se aplican procedimientos adecuados de normaliza-

ción y estandarización de intensidades, como los implementados en el pipeline de preprocesamiento descrito en la sección de Materiales y Métodos.

Por su parte, las imágenes correspondientes a la categoría MCI (Figura 7.3) revelan un patrón intermedio caracterizado por una marcada heterogeneidad en la distribución espacial de intensidades. Se identifican regiones con captación moderada del trazador, correspondientes a los tonos amarillo-naranjas, que coexisten con áreas de baja retención, lo que sugiere una deposición amiloide focal o en estadios iniciales. Esta variabilidad visual refleja la naturaleza clínicamente heterogénea del deterioro cognitivo leve como entidad transicional, en la cual algunos individuos presentan positividad amiloide y progresión hacia demencia, mientras que otros permanecen estables o incluso revierten a cognición normal [36].

Desde una perspectiva de modelado, esta heterogeneidad intraclase convierte a la categoría MCI en la más desafiante para los algoritmos de clasificación, ya que exige la identificación de patrones sutiles y no uniformes.

7.2. Evaluación cualitativa del preprocesamiento

Tras la aplicación del preprocesamiento, se realizó un análisis de la calidad de las imágenes resultantes con el objetivo de verificar la efectividad de cada etapa del procesamiento y evaluar la consistencia espacial y de intensidades a través de las diferentes modalidades de PET y categorías diagnósticas.

7.2.1. Análisis visual del preprocesamiento por grupo y categoría

Con el fin de verificar la correcta aplicación y consistencia del preprocesamiento, se realizó un análisis visual de las imágenes resultantes para cada combinación de categoría diagnóstica (AD, CN y MCI), grupo funcional del trazador (Amiloide, Metabólico y Tau) y subconjunto de datos (entrenamiento, validación y prueba). Las Figuras 7.5 a 7.12 muestran ejemplos representativos de dichas combinaciones.

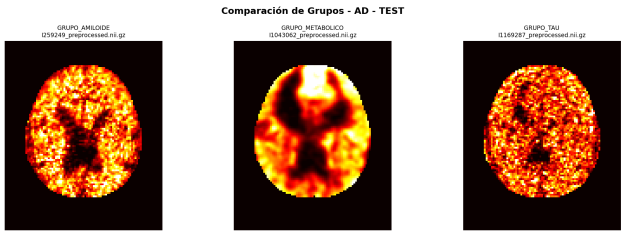


Figura 7.4: Imágenes preprocesadas de la categoría AD en el conjunto de prueba (test). De izquierda a derecha: Grupo Amiloide, Grupo Metabólico (FDG) y Grupo Tau.

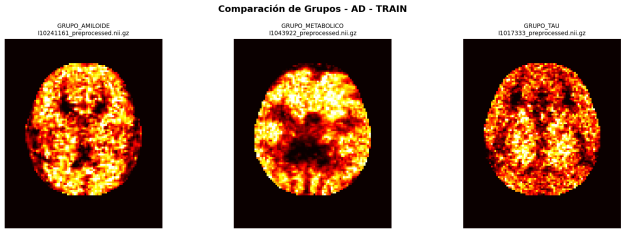


Figura 7.5: Imágenes preprocesadas de la categoría AD en el conjunto de entrenamiento (train).

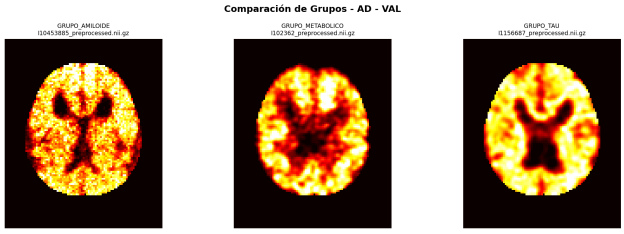


Figura 7.6: Imágenes preprocesadas de la categoría AD en el conjunto de validación (val).

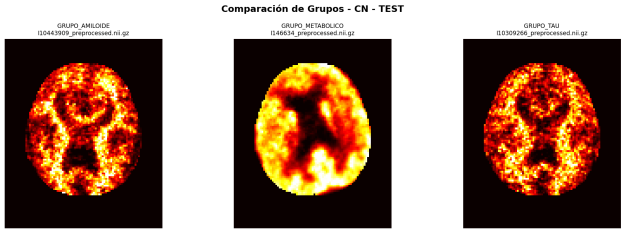


Figura 7.7: Imágenes preprocesadas de la categoría CN en el conjunto de prueba.

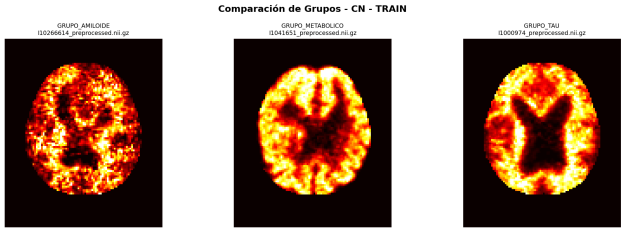


Figura 7.8: Imágenes preprocesadas de la categoría CN en el conjunto de entrenamiento.

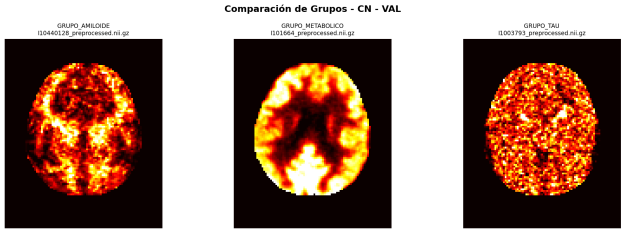


Figura 7.9: Imágenes preprocesadas de la categoría CN en el conjunto de validación.

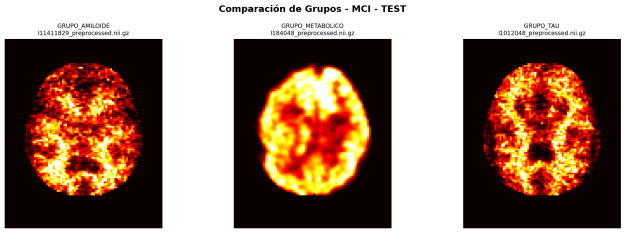


Figura 7.10: Imágenes preprocesadas de la categoría MCI en el conjunto de prueba.

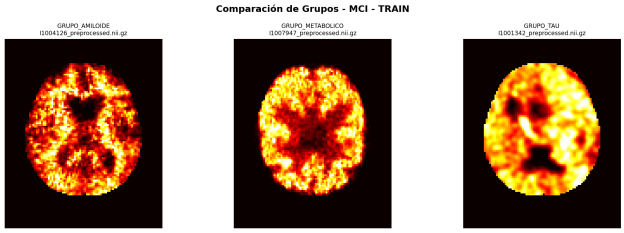


Figura 7.11: Imágenes preprocesadas de la categoría MCI en el conjunto de entrenamiento.

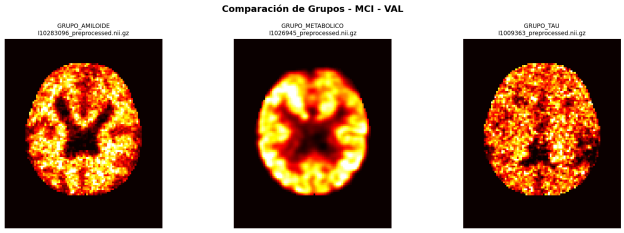


Figura 7.12: Imágenes preprocesadas de la categoría MCI en el conjunto de validación.

Este análisis permitió evaluar cualitativamente el efecto de las etapas de normalización de intensidades, alineación espacial y recorte de volumen, así como confirmar la preservación de patrones anatómicos y metabólicos relevantes tras el preprocesamiento. Asimismo, la inspección visual facilitó

la identificación de posibles artefactos, inconsistencias entre subconjuntos o variaciones sistemáticas entre grupos de trazadores que pudieran afectar el entrenamiento y la generalización de los modelos de aprendizaje profundo.

La coherencia observada entre las imágenes pertenecientes a diferentes particiones del conjunto de datos respalda la correcta separación entre los conjuntos de entrenamiento, validación y prueba, y sugiere que el preprocesamiento se aplicó de manera uniforme y reproducible en todo el conjunto de datos.

7.2.2. Verificación de la calidad del preprocesamiento

El análisis visual sistemático de las imágenes preprocesadas permitió verificar la correcta ejecución y efectividad de cada uno de los componentes del preprocesamiento implementado, asegurando la coherencia espacial, la estandarización de intensidades y la preservación de la información anatómica relevante. Es importante mencionar que, durante el proceso de preprocesamiento y verificación de la calidad de las imágenes, se identificaron 61 estudios PET que no cumplían con los criterios mínimos de calidad requeridos para su inclusión en el análisis, debido a inconsistencias en el registro, artefactos visuales o pérdida de información relevante tras las etapas de normalización y enmascaramiento. De esta manera, la base de datos final quedó consolidada con un total de 5,612 estudios PET.

7.2.2.1. Alineación espacial (registro rígido)

La correspondencia anatómica precisa entre las tres modalidades PET (Amiloide, Metabólico y Tau) dentro de cada fila de las Figuras 7.4 a 7.12 confirma la efectividad del registro rígido al atlas de referencia. Estructuras anatómicas clave, como los ventrículos laterales, identificables como regiones centrales hipointensas, ocupan posiciones espaciales idénticas a través de las distintas modalidades, lo que evidencia una correcta convergencia del algoritmo de optimización basado en información mutua de Mattes.

La ausencia de desalineamientos visibles resulta especialmente relevante considerando que las imágenes PET provienen de adquisiciones independientes, potencialmente realizadas en diferentes sesiones y con variaciones en el posicionamiento del sujeto. En este contexto, la consistencia espacial observada indica que el registro rígido eliminó eficazmente la variabilidad geométrica, una condición indispensable para que los modelos de aprendizaje profundo identifiquen patrones anatómicamente coherentes.

7.2.2.2. Normalización de intensidades (percentiles 1-99)

La normalización de intensidades al rango [0,1], basada en percentiles robustos (p1, p99), permitió estandarizar el rango dinámico de las imágenes sin comprometer las características diagnósticas relevantes. Esta efectividad se evidencia en varios aspectos:

- No se observan regiones saturadas ni artefactos visuales asociados a intensidades atípicas, lo que indica que el recorte por percentiles eliminó adecuadamente los valores extremos sin

pérdida de información significativa.

- Las imágenes correspondientes a una misma modalidad y categoría diagnóstica presentan distribuciones de intensidad visualmente homogéneas. Por ejemplo, las imágenes de amiloide CN mantienen tonos predominantemente rojizos-naranjas, mientras que las imágenes de amiloide AD exhiben una captación elevada de tonos amarillo-blancos, sin variaciones abruptas atribuibles a fallos de normalización.

- Se conserva una diferenciación clara entre la corteza cerebral, la sustancia blanca subcortical y los espacios ventriculares, lo que confirma que la normalización no introdujo una homogeneización artificial que pudiera degradar la información estructural relevante.

7.2.2.3. Enmascaramiento cerebral

La aplicación de la máscara cerebral binaria logró una eliminación efectiva de tejidos extracerebrales, preservando íntegramente el parénquima cerebral:

- En todas las imágenes, el fondo aparece completamente negro fuera del contorno cerebral, sin restos de señal provenientes del cuero cabelludo, hueso craneal u otros tejidos no relevantes. Esta supresión es esencial para evitar que los modelos aprendan características espurias ajenas al cerebro.

- Los márgenes corticales externos se mantienen intactos, sin signos de erosión artificial derivados de un enmascaramiento excesivamente restrictivo. Esta preservación es crítica, dado que la patología amiloide y tau se manifiesta principalmente a nivel cortical.

- Las estructuras subcorticales, como los ganglios basales y el tálamo, permanecen claramente visibles, especialmente en las imágenes FDG, tal como se evidencia en las Figuras 7.9 y 7.12, confirmando que la máscara no excluyó inadvertidamente regiones cerebrales profundas de interés funcional.

7.3. Evaluación de los modelos implementados

7.3.1. Evaluación de los modelos de redes neuronales convolucionales (CNN)

7.3.1.1. Evaluación del modelo ResNet3D

1. Análisis de las curvas de aprendizaje: La Figura 7.13 presenta las curvas de exactitud (*accuracy*) en función del número de épocas de entrenamiento para los conjuntos de entrenamiento y validación.

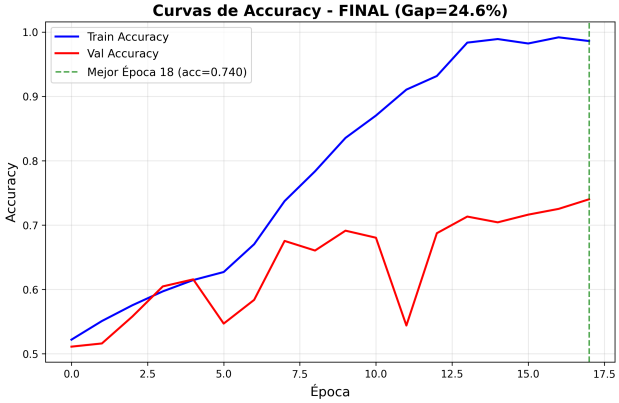


Figura 7.13: Curvas de exactitud del modelo ResNet3D durante el entrenamiento.

El análisis de las curvas de entrenamiento y validación permite evaluar la dinámica de aprendizaje y la capacidad de generalización del modelo. En la fase de entrenamiento, la exactitud muestra un incremento monotónico y sostenido, pasando de valores cercanos a 0.52 hasta aproximadamente 0.985 en las épocas finales. Este comportamiento evidencia una convergencia efectiva del proceso de optimización y confirma la capacidad de la arquitectura ResNet3D para aprender representaciones altamente discriminativas a partir de los volúmenes PET preprocesados, sin indicios de inestabilidad o estancamiento temprano.

En contraste, la curva de validación presenta una mayor variabilidad a lo largo del entrenamiento, con fluctuaciones pronunciadas que incluyen descensos temporales del desempeño, particularmente entre las épocas 8 y 12. Esta inestabilidad sugiere una sensibilidad del modelo a la composición del conjunto de validación y puede atribuirse a la presencia de muestras visualmente ambiguas, especialmente dentro de la categoría MCI, caracterizada por una elevada heterogeneidad clínica.

El máximo desempeño en validación se alcanzó en la época 18, con una exactitud de 0.740, identificándose este punto como el mejor compromiso entre aprendizaje y generalización. En consecuencia, se seleccionó este modelo como óptimo mediante un criterio de *early stopping* basado en validación. A partir de esta época, el incremento continuo de la exactitud en entrenamiento no se traduce en mejoras sostenidas en validación, lo que sugiere un inicio de sobreajuste.

Finalmente, la brecha de generalización entre las curvas de entrenamiento y validación alcanza un valor aproximado del 24.6 % en la última época. Si bien esta diferencia puede interpretarse como indicativa de un sobreajuste moderado, su magnitud debe analizarse considerando el desbalance de clases (con una razón aproximada de 5.2:1 entre CN y AD) y el tamaño limitado del conjunto de validación, factores que pueden amplificar las fluctuaciones observadas. No obstante, la evaluación definitiva de la capacidad de generalización del modelo se sustenta en el desempeño obtenido sobre el conjunto de prueba independiente, el cual proporciona una estimación más robusta y clínicamente relevante del comportamiento esperado del sistema.

2. Desempeño en el conjunto de validación: Los resultados cuantitativos del modelo ResNet3D en el conjunto de validación se presentan a continuación, desglosados por clase diagnóstica y complementados con promedios macro y micro que proporcionan una visión global del desempeño clasificatorio.

El modelo alcanzó una **exactitud global de 0.7400** en el conjunto de validación, lo que indica que aproximadamente tres de cada cuatro muestras fueron clasificadas correctamente. Este nivel de desempeño resulta alentador considerando la complejidad inherente del problema de clasificación de tres clases y el pronunciado desbalanceo del conjunto de datos.

Desempeño por clase:

El Cuadro 7.1 muestra el desempeño de cada clase después de la evaluación en el conjunto de validación.

Métrica	**AD**	**CN**	**MCI**
Precisión	0.6705	0.7627	0.7225
Recall	0.6082	0.8207	0.6675
Especificidad	0.9680	0.7332	0.8344
F1-Score	0.6378	0.7906	0.6939
AUC	0.8897	0.8452	0.8134

Cuadro 7.1: Comparación de métricas de desempeño por clase

El análisis del desempeño por clase diagnóstica evidencia comportamientos diferenciados entre las categorías evaluadas. En primer lugar, la clase **AD** presenta un balance adecuado entre sensibilidad y precisión. En particular, el *recall* de 0.6082 indica que el modelo identifica correctamente cerca del 61 % de los casos de Alzheimer en el conjunto de validación. Adicionalmente, la elevada especificidad (0.9680) demuestra una alta capacidad para descartar sujetos no-AD, reduciendo de manera significativa la tasa de falsos positivos. En conjunto, el valor de AUC de 0.8897 respalda una capacidad discriminativa robusta del modelo para diferenciar esta clase frente a las demás.

Por su parte, la clase **CN** exhibe el mejor desempeño global del modelo, reflejado en un F1-score de 0.7906. El *recall* de 0.8207 evidencia que más del 82 % de los controles normales son correctamente identificados, mientras que la precisión de 0.7627 indica una tasa de acierto elevada en las predicciones de esta categoría. Este comportamiento superior puede explicarse, en parte, por la mayor representación de la clase CN en el conjunto de entrenamiento (50.9 %), lo cual facilita el aprendizaje de patrones característicos de individuos cognitivamente sanos.

Finalmente, la clase **MCI** presenta un desempeño intermedio, con un F1-score de 0.6939. El *recall* de 0.6675 sugiere que aproximadamente dos tercios de los casos MCI son detectados correctamente, mientras que la especificidad de 0.8344 indica una capacidad aceptable para evitar confusiones con otras categorías diagnósticas. Este resultado es consistente con la naturaleza transicional y heterogénea del MCI, la cual representa un reto inherente para los sistemas de clasificación automática, tal como se evidenció en el análisis exploratorio previo.

Promedios agregados:

- Promedios Macro: Precisión = 0.7186, Recall = 0.6988, F1-Score = 0.7075

- Promedios Micro: Precisión = Recall = F1-Score = 0.7400

Los promedios macro, que otorgan igual peso a cada clase independientemente de su representación en el conjunto de datos, muestran valores ligeramente inferiores a los promedios micro, los cuales coinciden con la exactitud global. Esta diferencia sugiere que el desempeño del modelo es ligeramente superior en las clases mayoritarias, un patrón esperado en contextos de desbalance de clases.

3. Análisis de la matriz de confusión en validación: La matriz de confusión presentada en la Figura 7.14 proporciona una visualización detallada de los patrones de clasificación y error del modelo ResNet3D en el conjunto de validación.

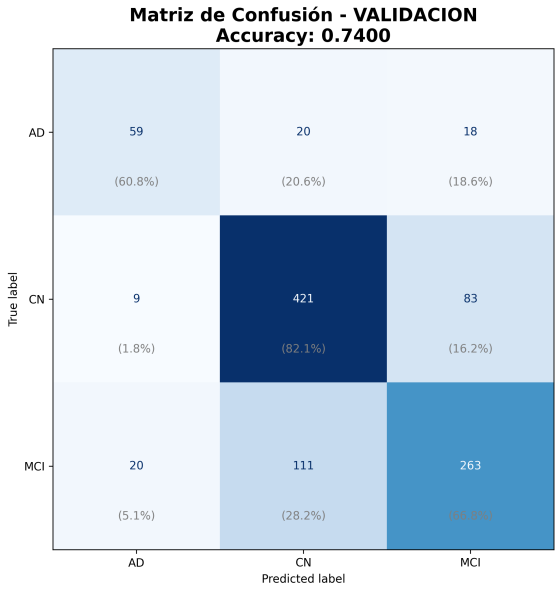


Figura 7.14: Matriz de confusión del modelo ResNet3D en el conjunto de validación.

Estos valores confirman que la clase CN presenta la mayor tasa de clasificación correcta, seguida por MCI y, finalmente, AD, un patrón consistente con las métricas de recall reportadas previamente.

4. Desempeño en el conjunto de prueba La evaluación en el conjunto de prueba proporciona la estimación más rigurosa y objetiva de la capacidad de generalización del modelo ResNet3D, al considerar datos completamente no vistos durante las etapas de entrenamiento y selección de hiperparámetros. En este sentido, los resultados obtenidos permiten valorar de manera confiable el comportamiento del modelo en escenarios de aplicación clínica real.

El modelo alcanzó una **exactitud global de 0.7034** en el conjunto de prueba, lo que representa una disminución marginal de apenas **3.67 % con respecto al desempeño en validación** (0.7400). Esta diferencia reducida constituye un indicador sólido de estabilidad y generalización, sugiriendo

que el modelo no incurrió en sobreajuste al conjunto de validación y que conserva su capacidad discriminativa al enfrentarse a datos previamente no observados.

Desempeño por clase en el conjunto de prueba: El Cuadro 7.2 muestra el desempeño de cada clase después de la evaluación en el conjunto de prueba.

Cuadro 7.2: Desempeño por clase en el conjunto de prueba

Métrica	AD	CN	MCI
Precisión	0.8000 (+0.1295)	0.7458 (−0.0169)	0.6327 (−0.0898)
Recall	0.5714 (−0.0368)	0.7458 (−0.0749)	0.6889 (+0.0214)
Especificidad	0.9808 (+0.0128)	0.7458 (+0.0126)	0.7534 (−0.0810)
F1-Score	0.6667 (+0.0289)	0.7458 (−0.0448)	0.6596 (−0.0343)
AUC	0.9581 (+0.0684)	0.8219 (−0.0233)	0.7976 (−0.0158)

El análisis del desempeño en el conjunto de prueba confirma la capacidad de generalización del modelo, evidenciando patrones diferenciados entre las clases diagnósticas. En la clase **AD**, se observa una mejora sustancial en la precisión y, especialmente, en el área bajo la curva ROC, alcanzando un valor de AUC de 0.9581, lo que corresponde a una capacidad discriminativa excelente. Asimismo, la especificidad extremadamente alta (0.9808) indica que el modelo rara vez clasifica de manera errónea sujetos no AD como Alzheimer, un aspecto particularmente relevante desde la perspectiva clínica al minimizar el riesgo de sobrediagnóstico.

Sin embargo, el *recall* disminuye ligeramente hasta 0.5714, lo que implica que aproximadamente cuatro de cada siete casos de AD son identificados correctamente en datos no vistos. Este comportamiento sugiere un compromiso entre sensibilidad y especificidad, reflejando una estrategia conservadora del modelo, en la que se prioriza la certeza diagnóstica al asignar la categoría de AD, aun a costa de omitir algunos casos verdaderos.

Por su parte, la clase **CN** mantiene un desempeño relativamente estable en el conjunto de prueba, con métricas que se sitúan en el rango de 0.74 a 0.82. La reducción moderada del *recall* sugiere una leve pérdida de sensibilidad en la identificación de controles normales, posiblemente asociada a la presencia de sujetos con perfiles limítrofes cercanos a MCI. No obstante, el AUC de 0.8219 confirma una capacidad discriminativa adecuada para esta categoría.

Finalmente, la clase **MCI** continúa representando el mayor desafío para el modelo, en concordancia con su heterogeneidad clínica y biológica. Aunque se evidencia una disminución en la precisión, el *recall* muestra una ligera mejora, indicando una mayor capacidad del modelo para capturar casos MCI en el conjunto de prueba. En este contexto, el valor de AUC de 0.7976 sugiere una discriminación razonable, aunque inferior a la observada en las clases AD y CN, resultado coherente con la naturaleza transicional de esta condición.

Promedios agregados en el conjunto de prueba:

- Promedios macro: Precisión = 0.7261, Recall = 0.6687, F1-Score = 0.6907
- Promedios micro: Precisión = Recall = F1-Score = 0.7034

Los promedios macro y micro presentan valores coherentes con los observados en el conjunto de validación, lo que confirma que el modelo mantiene un balance razonable entre clases a pesar del desbalanceo inherente del conjunto de datos. En conjunto, estos resultados respaldan la estabilidad del modelo ResNet3D y su potencial como herramienta de apoyo al diagnóstico basado en imágenes PET, particularmente para la identificación confiable de casos de enfermedad de Alzheimer.

5. Análisis de la matriz de confusión en el conjunto de prueba La matriz de confusión del conjunto de prueba (Figura 7.15) proporciona información complementaria sobre los patrones de error en datos completamente independientes.

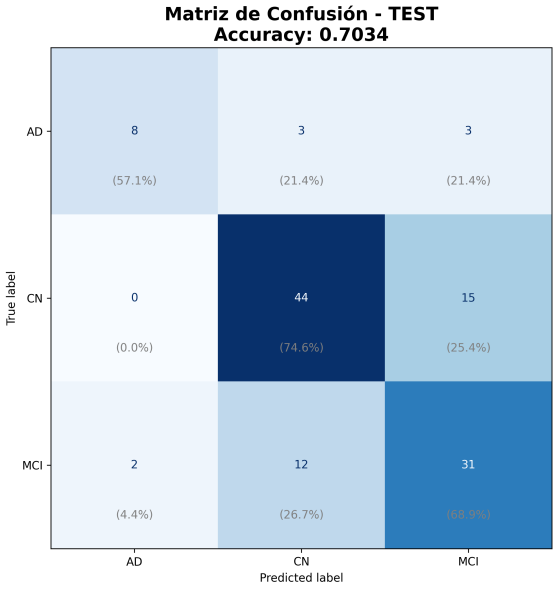


Figura 7.15: Matriz de confusión del modelo ResNet3D en el conjunto de prueba.

Clasificaciones correctas:

- 8 de 14 casos AD (57.1 %)
- 44 de 59 casos CN (74.6 %)
- 31 de 45 casos MCI (68.9 %)

El tamaño reducido del conjunto de prueba de 118 muestras totales introduce mayor variabilidad en las tasas de clasificación correcta por clase, particularmente evidente en AD, donde la muestra consta de solo 14 casos. No obstante, las proporciones generales mantienen el patrón observado en validación, con CN exhibiendo la mayor tasa de acierto.

6. Análisis de las curvas ROC: Curvas ROC en el conjunto de validación:

La Figura 7.16 presenta las curvas ROC para cada clase en el conjunto de validación, junto con sus correspondientes valores de AUC.

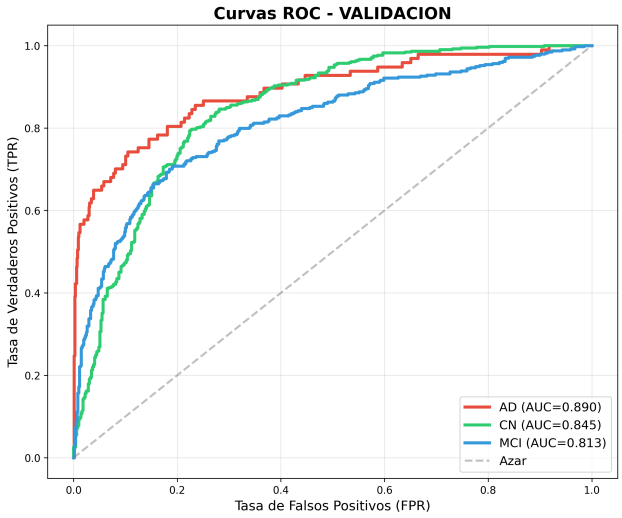


Figura 7.16: Curvas ROC del modelo ResNet3D en el conjunto de validación.

El análisis de las curvas ROC en validación revela:

- **AD (AUC=0.890):** La curva roja muestra una elevación pronunciada en la región de baja tasa de falsos positivos, alcanzando aproximadamente 65 % de sensibilidad con menos del 5 % de falsos positivos. Este comportamiento indica que el modelo puede configurarse para detectar la mayoría de los casos AD mientras mantiene una tasa de falsos positivos muy baja, propiedad altamente deseable en aplicaciones de screening clínico.

- **CN (AUC=0.845):** La curva verde presenta un ascenso sostenido y relativamente uniforme, reflejando un balance consistente entre sensibilidad y especificidad a través de diversos umbrales de decisión.

- **MCI (AUC=0.813):** La curva azul, aunque exhibe la menor AUC de las tres clases, mantiene una separación clara respecto al clasificador aleatorio, indicando capacidad discriminativa genuina a pesar de la heterogeneidad de esta categoría.

Curvas ROC en el conjunto de prueba:

La Figura 7.17 muestra las curvas ROC en el conjunto de prueba, permitiendo evaluar la estabilidad de la capacidad discriminativa en datos no vistos.

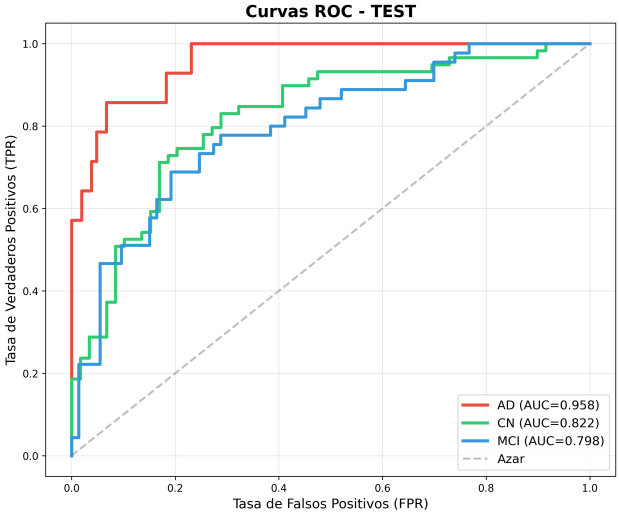


Figura 7.17: Curvas ROC del modelo ResNet3D en el conjunto de prueba.

Los resultados en el conjunto de prueba muestran características notables:

- **AD (AUC=0.958):** El incremento sustancial del AUC respecto a la validación de 0.890 a 0.958 representa un resultado excepcional, ubicando al modelo en el rango de discriminación excelente para esta clase. La curva roja muestra una ascensión casi vertical en la región inicial, alcanzando más del 70 % de sensibilidad con tasas de falsos positivos inferiores al 5 %.

- **CN (AUC=0.822) y MCI (AUC=0.798):** Ambas clases mantienen AUCs consistentes con la validación, con reducciones mínimas que confirman la estabilidad del desempeño discriminativo.

7.3.1.2. Evaluación del modelo de Transfer Learning con ResNet-10

1. Análisis de las curvas de aprendizaje: La implementación del *transfer learning* en este trabajo siguió una estrategia de entrenamiento en dos fases, diseñada para optimizar progresivamente diferentes niveles de la arquitectura:

Fase 1 - Ajuste de capas superiores: La Figura 7.18 ilustra el comportamiento del modelo durante esta primera fase.

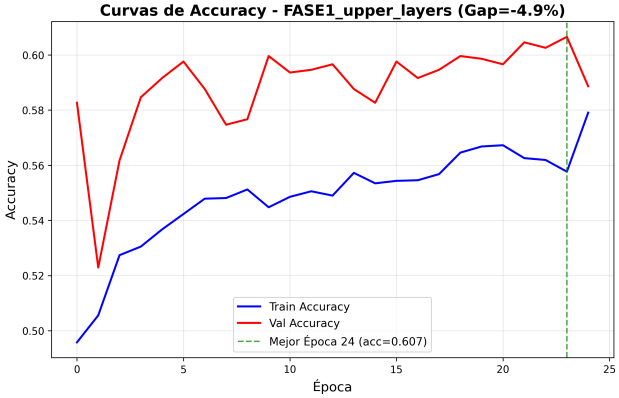


Figura 7.18: Curvas de exactitud del modelo Transfer Learning ResNet-10 durante la Fase 1 (ajuste de capas superiores).

El análisis de las curvas de entrenamiento y validación durante la Fase 1 evidencia un comportamiento característico del aprendizaje por *transfer learning*. En las primeras épocas, ambas curvas muestran un incremento pronunciado, alcanzando valores de exactitud cercanos a 0.55 hacia la época 5. Este desempeño inicial refleja el beneficio del conocimiento transferido, donde las representaciones preentrenadas proporcionan un punto de partida superior en comparación con una inicialización aleatoria.

Posteriormente, a partir de la época 10, se observa una estabilización relativa en ambas curvas, acompañada de fluctuaciones moderadas. Este comportamiento sugiere que el ajuste exclusivo de las capas superiores alcanza rápidamente un límite en la capacidad representacional del modelo. En consecuencia, la exactitud de entrenamiento se mantiene en un rango de 0.54 a 0.58, valores considerablemente inferiores a los obtenidos durante el entrenamiento completo de la arquitectura ResNet3D.

Un aspecto particularmente relevante es que la curva de validación mantiene valores iguales o superiores a los de entrenamiento durante la mayor parte del proceso, resultando en una brecha de generalización negativa del orden del $-4{,}9\,\%$. Este fenómeno, contrario al sobreajuste tradicional, es indicativo de un régimen de *underfitting*, en el cual la capacidad del modelo no es plenamente explotada.

Finalmente, el mejor desempeño en validación, correspondiente a una exactitud de 0.607 alcanzada en la época 24, resulta inferior al obtenido por ResNet3D en su configuración óptima. En conjunto, estos resultados confirman la hipótesis de que las capas convolucionales inferiores, congeladas durante esta fase, contienen representaciones generales que requieren una adaptación específica al dominio de imágenes PET amiloide para capturar de manera adecuada la complejidad del problema de clasificación.

Fase 2 - Ajuste fino completo: Basándose en los resultados de la Fase 1, se procedió a descongelar todas las capas de la red, permitiendo el ajuste fino de la arquitectura completa con una tasa de aprendizaje reducida. La Figura 7.19 presenta las curvas de aprendizaje correspondientes a esta segunda fase.

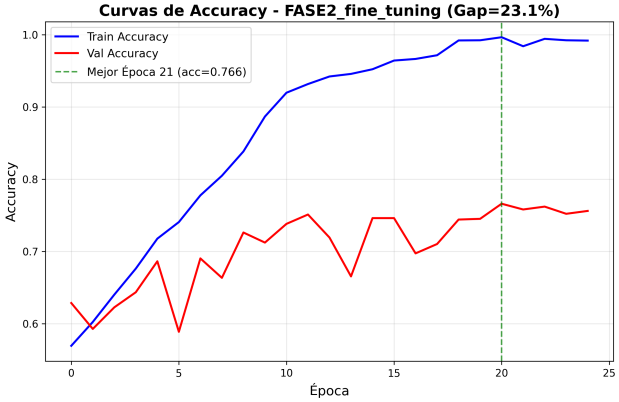


Figura 7.19: Curvas de exactitud del modelo Transfer Learning ResNet-10 durante la Fase 2 (ajuste fino completo).

El comportamiento observado durante la **Fase 2** difiere sustancialmente del evidenciado en la Fase 1, poniendo de manifiesto los efectos del ajuste fino completo de la arquitectura. En particular, la curva de entrenamiento presenta un ascenso sostenido y pronunciado, alcanzando valores cercanos a 0.99 en las épocas finales. Este incremento, partiendo de un valor inicial aproximado de 0.57 heredado de la Fase 1, evidencia que el descongelamiento de las capas convolucionales inferiores libera una capacidad representacional significativa, permitiendo al modelo capturar patrones más complejos del dominio PET amiloide.

Por su parte, la curva de validación muestra un incremento más moderado, iniciando alrededor de 0.63 y alcanzando un máximo de 0.766 en la época 21. Si bien este valor representa una mejora sustancial de aproximadamente 15.9 puntos porcentuales con respecto a la Fase 1, el crecimiento es menos pronunciado que en la curva de entrenamiento y se encuentra acompañado de fluctuaciones considerables a lo largo del proceso.

A diferencia de la Fase 1, la Fase 2 desarrolla una brecha de generalización creciente entre las curvas de entrenamiento y validación, alcanzando un valor cercano al 23.1 % al final del entrenamiento. Este comportamiento sugiere que el ajuste fino completo permite una optimización intensa sobre el conjunto de entrenamiento, aunque con una traducción limitada hacia mejoras sostenidas en la validación. Asimismo, la elevada variabilidad observada en la curva de validación, particularmente entre las épocas 10 y 18, indica una sensibilidad del modelo a la composición de los *mini-batches* y una posible inestabilidad en las regiones del espacio de parámetros exploradas durante el ajuste fino.

En conjunto, la estrategia de entrenamiento en dos fases resulta efectiva para mejorar el desempeño global del modelo en comparación con el ajuste exclusivo de las capas superiores. En particular, se alcanza una exactitud de validación de 0.766, superando en 2.6 puntos porcentuales el mejor resultado obtenido por ResNet3D entrenado desde cero (0.740). Estos hallazgos sugieren que la combinación de conocimiento transferido y ajuste fino completo aporta ventajas claras para el problema de clasificación abordado.

2. Desempeño en el conjunto de validación: Los resultados cuantitativos del modelo basado en *transfer learning* evaluado en el conjunto de validación, correspondientes a la Fase 2 y a la época 21 (época de mejor desempeño), se presentan a continuación, organizados por categoría diagnóstica. Este análisis permite cuantificar de manera objetiva el impacto de la estrategia de *transfer learning* en comparación con el modelo base ResNet3D entrenado desde cero.

El modelo alcanzó una **exactitud global de 0.7659** en el conjunto de validación, lo que representa una mejora absoluta de **2.59 puntos porcentuales** respecto al desempeño obtenido por ResNet3D (0.7400). Este incremento, aunque moderado, es consistente y clínicamente relevante, indicando que la inicialización del modelo con pesos preentrenados facilitó un aprendizaje más eficiente y una mejor capacidad de generalización.

Desempeño por clase diagnóstica: El Cuadro 7.3 muestra el desempeño de cada clase después de la evaluación en el conjunto de validación.

Cuadro 7.3: Desempeño por clase diagnóstica

Métrica	**AD**	**CN**	**MCI**
Precisión	0.7024 (+0.0319)	0.7754 (+0.0127)	0.7657 (+0.0432)
Recall	0.6082 (0.0000)	0.8616 (+0.0409)	0.6802 (+0.0127)
Especificidad	0.9724 (+0.0044)	0.7393 (+0.0061)	0.8656 (+0.0312)
F1-Score	0.6519 (+0.0141)	0.8163 (+0.0257)	0.7204 (+0.0265)
AUC	0.9001 (+0.0104)	0.8868 (+0.0416)	0.8571 (+0.0437)

Valores entre paréntesis: diferencia respecto a ResNet3D

El análisis comparativo del desempeño por clase evidencia mejoras consistentes del modelo propuesto frente a ResNet3D. En la clase **AD**, se observan incrementos en todas las métricas evaluadas, con excepción del *recall*, que permanece inalterado. La precisión alcanzada (0.7024) indica que aproximadamente siete de cada diez predicciones de Alzheimer son correctas, lo que contribuye a una reducción de la tasa de falsos positivos respecto al modelo base. Asimismo, la especificidad excepcionalmente alta (0.9724) confirma la capacidad del modelo para descartar adecuadamente casos no-AD, un aspecto de particular relevancia clínica para minimizar el riesgo de sobrediagnóstico.

Por su parte, la clase **CN** presenta el mejor desempeño absoluto entre todas las categorías, así como las mejoras más pronunciadas en comparación con ResNet3D. El *recall* de 0.8616 evidencia que más del 86 % de los sujetos cognitivamente normales son identificados correctamente, superando en más de cuatro puntos porcentuales al modelo base. Adicionalmente, el F1-score de 0.8163 constituye el valor más alto alcanzado por cualquier combinación de modelo y clase hasta este punto del análisis, reflejando un balance óptimo entre precisión y sensibilidad.

Finalmente, la clase **MCI**, tradicionalmente la más desafiante debido a su heterogeneidad clínica, muestra mejoras sustanciales en todas las métricas consideradas. Destacan especialmente los incrementos en precisión (+4.3 puntos porcentuales) y en AUC (+4.4 puntos porcentuales), lo que sugiere una mayor capacidad del modelo para discriminar correctamente esta categoría intermedia. En conjunto, el F1-score de 0.7204 refleja un equilibrio mejorado entre sensibilidad y precisión para esta población frontera.

Promedios agregados:

- Promedios macro: Precisión = 0.7478, Recall = 0.7167, F1-Score = 0.7295
- Promedios micro: Precisión = Recall = F1-Score = 0.7659

Los promedios macro presentan incrementos consistentes respecto al modelo ResNet3D (precisión +0.0292, recall +0.0179 y F1-score +0.0220), confirmando que los beneficios del *transfer learning* no se concentran en una sola categoría, sino que se distribuyen de manera equilibrada entre las tres clases diagnósticas, incluso en presencia de desbalanceo del conjunto de datos.

3. Análisis de la matriz de confusión en validación: La matriz de confusión del modelo de *transfer learning* en validación (Figura 7.20) permite analizar los patrones específicos de clasificación y error.

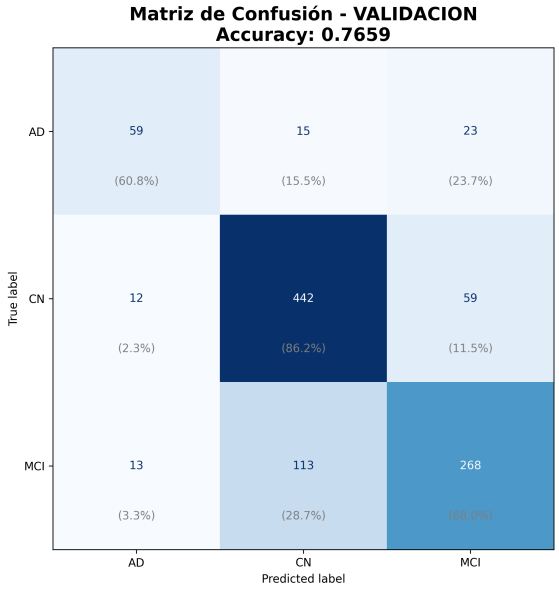


Figura 7.20: Matriz de confusión del modelo Transfer Learning ResNet-10 en el conjunto de validación.

Análisis diagonal (clasificaciones correctas):

- 59 de 97 casos AD (60.8 %) — idéntico a ResNet3D
- 442 de 513 casos CN (86.2 %) — mejora de +21 casos respecto a ResNet3D
- 268 de 394 casos MCI (68.0 %) — mejora de +5 casos respecto a ResNet3D

La mejora más sustancial se observa en la clase CN, donde 21 casos adicionales son clasificados correctamente, elevando la tasa de acierto del 82.1 % al 86.2 %.

4. Desempeño en el conjunto de prueba: La evaluación en el conjunto de prueba proporciona la medida definitiva de la capacidad de generalización del modelo de *transfer learning*.

El modelo alcanzó una **exactitud global de 0.7034** en el conjunto de prueba, **idéntica al desempeño de ResNet3D** en este conjunto. La diferencia entre validación y prueba es de **6.25 %**, ligeramente superior a la observada en ResNet3D (3.67 %), aunque permanece dentro del rango considerado como buena generalización.

Desempeño por clase en el conjunto de prueba: El Cuadro 7.4 muestra el desempeño de cada clase después de la evaluación en el conjunto de prueba.

Cuadro 7.4: Desempeño por clase en el conjunto de prueba

Métrica	**AD**	**CN**	**MCI**
Precisión	0.6923 (−0.1077)	0.7667 (+0.0209)	0.6222 (−0.0105)
Recall	0.6429 (+0.0715)	0.7797 (+0.0339)	0.6222 (−0.0667)
Especificidad	0.9615 (−0.0193)	0.7627 (+0.0169)	0.7671 (+0.0137)
F1-Score	0.6667 (0.0000)	0.7731 (+0.0273)	0.6222 (−0.0374)
AUC	0.8860 (−0.0721)	0.8779 (+0.0560)	0.8511 (+0.0535)

Valores entre paréntesis: diferencia respecto a ResNet3D en test

El análisis comparativo del desempeño por clase en el conjunto de prueba evidencia comportamientos diferenciados entre el modelo basado en *transfer learning* y ResNet3D. En la clase **AD**, se observa un *trade-off* respecto al modelo base, caracterizado por un incremento en la sensibilidad (64.3 % frente a 57.1 %) a costa de una reducción en la precisión y en el AUC. En términos absolutos, el modelo con *transfer learning* identifica correctamente 9 de los 14 casos AD, uno más que ResNet3D, aunque este beneficio se acompaña de un aumento en la tasa de falsos positivos.

Por su parte, la clase **CN** exhibe mejoras consistentes en todas las métricas evaluadas en comparación con ResNet3D, lo que confirma la ventaja del *transfer learning* para la identificación de sujetos cognitivamente normales. En este caso, el modelo clasifica correctamente 46 de los 59 casos CN, lo que representa dos aciertos adicionales respecto al modelo entrenado desde cero.

Finalmente, la clase **MCI** presenta resultados mixtos. Si bien se observa una disminución en el *recall* y en el F1-score, se registra un incremento en el AUC, lo que sugiere una mejora en la capacidad discriminativa global pese a una menor tasa de detección de casos. En términos absolutos, el modelo identifica correctamente 28 de los 45 casos MCI, tres menos que ResNet3D, reflejando la dificultad inherente a la clasificación de esta categoría intermedia.

Promedios agregados en prueba:

- Promedios Macro: Precisión = 0.6937, Recall = 0.6816, F1-Score = 0.6873
- Promedios Micro: Precisión = Recall = F1-Score = 0.7034

Los promedios macro son ligeramente inferiores a ResNet3D en prueba (precisión −0.0324, recall +0.0129, F1-score −0.0034), sugiriendo que las ventajas del *transfer learning* observadas en validación no se transfieren completamente al conjunto de prueba.

5. Análisis de la matriz de confusión en el conjunto de prueba: La matriz de confusión en el conjunto de prueba (Figura 7.21) revela los patrones de clasificación en datos completamente independientes.

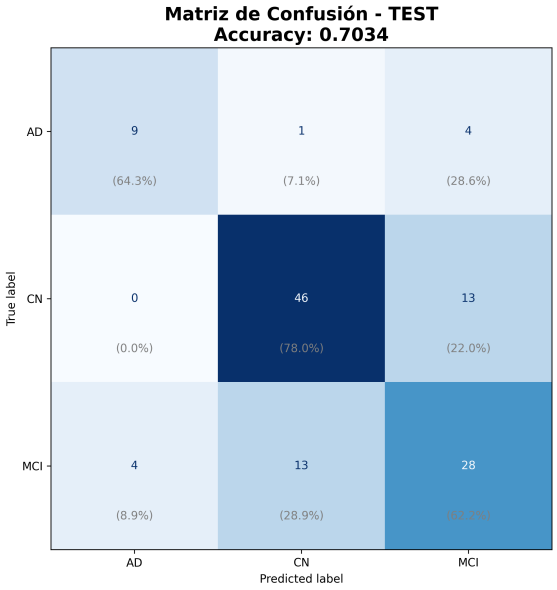


Figura 7.21: Matriz de confusión del modelo Transfer Learning ResNet-10 en el conjunto de prueba.

Clasificaciones correctas:

- 9 de 14 casos AD (64.3 %) — +1 caso respecto a ResNet3D
- 46 de 59 casos CN (78.0 %) — +2 casos respecto a ResNet3D
- 28 de 45 casos MCI (62.2 %) — −3 casos respecto a ResNet3D

6. Análisis de las curvas ROC Curvas ROC en el conjunto de validación:

La Figura 7.22 presenta las curvas ROC para cada clase en validación.

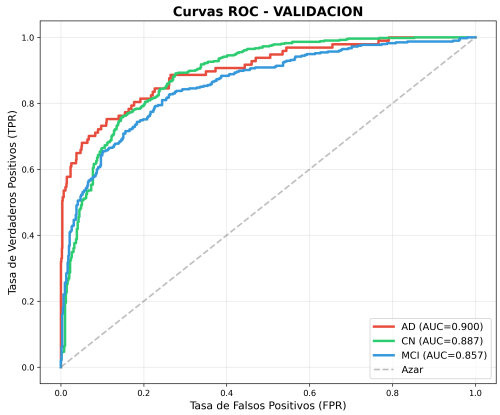


Figura 7.22: Curvas ROC del modelo Transfer Learning ResNet-10 en el conjunto de validación.

El análisis comparativo con ResNet3D revela:

- **AD (AUC=0.900 vs 0.890):** Mejora marginal de +0.010, con la curva manteniendo una elevación pronunciada en regiones de baja tasa de falsos positivos.
- **CN (AUC=0.887 vs 0.845):** Mejora sustancial de +0.042, representando el incremento más significativo entre las tres clases. La curva verde muestra una separación más clara respecto al clasificador aleatorio en todo el rango de umbrales.
- **MCI (AUC=0.857 vs 0.813):** Mejora notable de +0.044, elevando la capacidad discriminativa para esta categoría desafiante desde un nivel bueno hacia excelente.

Curvas ROC en el conjunto de prueba:
La Figura 7.23 muestra las curvas ROC en el conjunto de prueba.

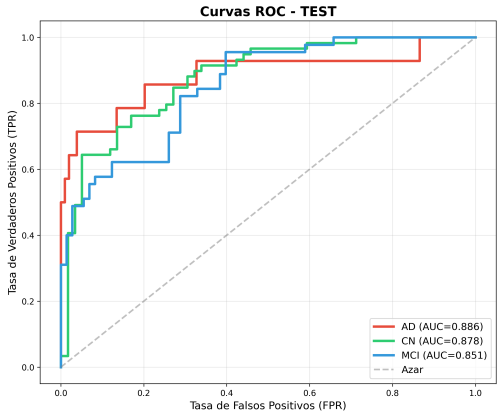


Figura 7.23: Curvas ROC del modelo Transfer Learning ResNet-10 en el conjunto de prueba.

Los resultados en la prueba muestran:

- **AD (AUC=0.886 vs 0.958 de ResNet3D:** Reducción significativa de −0.072, sugiriendo que las ventajas del *transfer learning* para AD no se generalizan al conjunto de prueba con la misma efectividad que ResNet3D.
- **CN (AUC=0.878 vs 0.822 de ResNet3D):** Mejora de +0.056, confirmando la ventaja consistente del *transfer learning* para la identificación de controles normales.
- **MCI (AUC=0.851 vs 0.798 de ResNet3D):** Mejora de +0.053, indicando una mejor capacidad discriminativa para esta categoría en datos no vistos.

La consistencia entre los AUCs de validación y prueba, con diferencias menores a 0.02 en todas las clases, confirma la estabilidad de la capacidad discriminativa del modelo de *transfer learning*.

7.3.1.3. Evaluación del modelo VoxCNN 3D

1. Análisis de las curvas de aprendizaje: La Figura 7.24 presenta las curvas de exactitud durante el entrenamiento del modelo VoxCNN 3D.

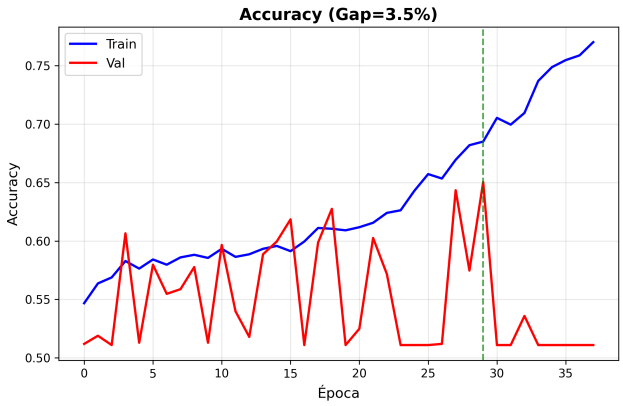


Figura 7.24: Curvas de exactitud del modelo VoxCNN 3D durante el entrenamiento.

El análisis de las curvas de aprendizaje del modelo **VoxCNN** revela un comportamiento marcadamente distinto al observado en ResNet3D y en el esquema de *transfer learning*, evidenciando limitaciones significativas en su capacidad de representación. En particular, la curva de entrenamiento muestra un ascenso gradual desde valores cercanos a 0.55 hasta estabilizarse alrededor de 0.78 hacia la época 38. Este incremento resulta considerablemente más modesto que el alcanzado por ResNet3D y por el modelo con ajuste fino completo, lo que indica que la arquitectura carece de la capacidad necesaria para ajustar plenamente el conjunto de entrenamiento.

En contraste, la curva de validación presenta fluctuaciones pronunciadas a lo largo de todo el proceso, oscilando en un rango estrecho entre aproximadamente 0.51 y 0.65, sin evidenciar una tendencia ascendente sostenida después de la época 25. Este comportamiento sugiere que el modelo alcanza rápidamente el límite de su capacidad de generalización, de modo que un entrenamiento adicional no se traduce en mejoras consistentes del desempeño.

Adicionalmente, las oscilaciones observadas en la curva de validación son más intensas que en los modelos previamente evaluados, con variaciones abruptas de hasta 10 puntos porcentuales entre épocas consecutivas. Esta inestabilidad puede atribuirse a la elevada sensibilidad de una arquitectura de capacidad limitada frente a la composición específica de los *mini-batches* empleados durante la validación.

Si bien VoxCNN presenta la menor brecha de generalización al final del entrenamiento, con una diferencia aproximada del 3.5 % entre las curvas de entrenamiento y validación, este resultado no debe interpretarse como evidencia de una generalización superior. Por el contrario, constituye un indicador de *underfitting* severo, en el cual el modelo no logra aprender adecuadamente ni siquiera los patrones presentes en el conjunto de entrenamiento. A diferencia de ResNet3D, donde se observa una divergencia creciente entre las curvas, VoxCNN mantiene ambas trayectorias relativamente próximas durante todo el proceso, confirmando que la arquitectura opera de manera sistemática por debajo de su capacidad óptima de representación.

En conjunto, este patrón de comportamiento sugiere que la estrategia de simplificación arquitectónica adoptada por VoxCNN resulta excesivamente restrictiva para el problema de clasificación abordado, sacrificando una porción significativa de capacidad expresiva en favor de la eficiencia computacional.

2. Desempeño en el conjunto de validación: Los resultados cuantitativos del modelo VoxCNN 3D en el conjunto de validación evidencian un desempeño significativamente inferior al de las arquitecturas ResNet3D y *transfer learning*.

El modelo alcanzó una **exactitud global de 0.6504** en el conjunto de validación, representando una reducción de **8.96 puntos porcentuales respecto a ResNet3D** (0.7400) y de **11.55 puntos porcentuales respecto al *transfer learning*** (0.7659). Este nivel de desempeño posiciona a VoxCNN como el modelo de peor rendimiento entre las tres arquitecturas CNN evaluadas.

Desempeño por clase: El Cuadro 7.5 muestra el desempeño de cada clase después de la evaluación en el conjunto de validación.

Cuadro 7.5: Desempeño por clase

Métrica	**AD**	**CN**	**MCI**
Precisión	0.5362 (−0.1343)	0.6718 (−0.0909)	0.6298 (−0.0927)
Recall	0.3814 (−0.2268)	0.8460 (+0.0253)	0.4619 (−0.2056)
Especificidad	0.9647 (−0.0033)	0.5682 (−0.1650)	0.8246 (−0.0098)
F1-Score	0.4458 (−0.1920)	0.7489 (−0.0417)	0.5329 (−0.1610)
AUC	0.8641 (−0.0256)	0.8101 (−0.0351)	0.7420 (−0.0714)

Valores entre paréntesis: diferencia respecto a ResNet3D

El análisis del desempeño por clase del modelo **VoxCNN** pone de manifiesto limitaciones sustanciales en su aplicabilidad para el problema abordado. En la clase **AD**, el desempeño resulta particularmente deficiente, con un *recall* de 0.3814, lo que implica que el modelo identifica correctamente menos de cuatro de cada diez casos de Alzheimer. Esta sensibilidad extremadamente baja es

clínicamente inaceptable, ya que aproximadamente el 62 % de los casos de AD no serían detectados. En concordancia, el F1-score de 0.4458 constituye el valor más bajo observado entre todas las clases y modelos evaluados, reflejando un balance pobre entre precisión y sensibilidad.

Por su parte, la clase **CN**, favorecida por su mayor representación en el conjunto de datos, exhibe el mejor desempeño relativo dentro de VoxCNN. El *recall* de 0.8460 resulta incluso ligeramente superior al alcanzado por ResNet3D; sin embargo, la precisión reducida (0.6718) evidencia una tasa elevada de falsos positivos. En consecuencia, el F1-score de 0.7489, aunque es el más alto obtenido por esta arquitectura, permanece por debajo de los valores alcanzados por los modelos previamente evaluados.

Finalmente, la clase **MCI** presenta métricas consistentemente bajas, con un *recall* de 0.4619, indicando que menos de la mitad de los casos son identificados correctamente. Asimismo, el valor de AUC de 0.7420 se sitúa en el límite inferior del rango considerado aceptable, lo que sugiere una capacidad discriminativa marginal para esta categoría. En conjunto, estos resultados confirman que la limitada capacidad representacional de VoxCNN afecta de manera crítica su desempeño, especialmente en las clases clínicamente más relevantes.

Promedios agregados:

- Promedios Macro: Precisión = 0.6126, Recall = 0.5631, F1-Score = 0.5759
- Promedios Micro: Precisión = Recall = F1-Score = 0.6504

Los promedios macro revelan un desempeño global deficiente, con valores sustancialmente inferiores a los de ResNet3D y *transfer learning* en todas las métricas. La diferencia de más de 10 puntos porcentuales en recall macro respecto a ResNet3D (0.5631 vs 0.6988) es particularmente preocupante desde una perspectiva clínica.

3. Análisis de la matriz de confusión en validación: La matriz de confusión del modelo VoxCNN en validación (Figura 7.25) revela patrones de error severos y sistemáticos.

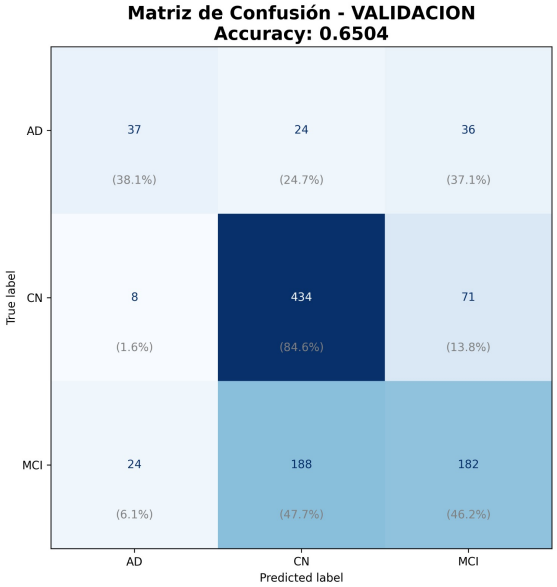


Figura 7.25: Matriz de confusión del modelo VoxCNN 3D en el conjunto de validación.

Análisis diagonal (clasificaciones correctas):

- 37 de 97 casos AD (38.1 %) — reducción de −22 casos respecto a ResNet3D
- 434 de 513 casos CN (84.6 %) — mejora de +13 casos respecto a ResNet3D
- 182 de 394 casos MCI (46.2 %) — reducción de −81 casos respecto a ResNet3D

Los resultados diagonales evidencian un desbalance severo en el desempeño entre clases. Mientras que CN alcanza una tasa de clasificación correcta superior incluso a ResNet3D, las clases AD y MCI sufren degradaciones dramáticas, particularmente MCI donde menos de la mitad de los casos son identificados correctamente.

4. Desempeño en el conjunto de prueba: La evaluación en el conjunto de prueba confirma y agudiza las limitaciones observadas en la validación, con VoxCNN exhibiendo el desempeño más deficiente entre todas las arquitecturas de CNN evaluadas.

El modelo alcanzó una **exactitud global de 0.5763** en el conjunto de prueba, representando una degradación adicional de **7.41 puntos porcentuales respecto a validación** (0.6504) y posicionándose **12.71 puntos porcentuales por debajo de ResNet3D** en este conjunto. Este nivel de desempeño se aproxima apenas al de un clasificador con capacidad discriminativa marginal.

Desempeño por clase en el conjunto de prueba: El Cuadro 7.6 muestra el desempeño de cada clase después de la evaluación en el conjunto de prueba.

Cuadro 7.6: Desempeño por clase en el conjunto de prueba

Métrica	AD	CN	MCI
Precisión	0.1250 (−0.6750)	0.6216 (−0.1242)	0.5833 (−0.0494)
Recall	0.0714 (−0.5000)	0.7797 (+0.0339)	0.4667 (−0.2222)
Especificidad	0.9327 (−0.0481)	0.5254 (−0.2204)	0.7945 (+0.0411)
F1-Score	0.0909 (−0.5758)	0.6917 (−0.0541)	0.5185 (−0.1411)
AUC	0.7273 (−0.2308)	0.7972 (−0.0247)	0.7766 (−0.0210)

Valores entre paréntesis: diferencia respecto a ResNet3D en test

El análisis del desempeño de **VoxCNN** en el conjunto de prueba revela un deterioro severo de la capacidad de generalización del modelo, particularmente en la identificación de la enfermedad de Alzheimer. En la clase **AD**, se observa un colapso catastrófico del desempeño, con un *recall* de apenas 0.0714, lo que implica que el modelo identifica correctamente solo 1 de los 14 casos AD (7.1 %), dejando inadvertidos los 13 restantes. Asimismo, la precisión de 0.1250 indica que únicamente una de cada ocho predicciones de AD es correcta. En consecuencia, el F1-score de 0.0909 resulta excepcionalmente bajo, evidenciando que VoxCNN carece prácticamente de capacidad operativa para la detección de Alzheimer en datos no vistos, lo cual lo hace clínicamente inviable.

Por su parte, la clase **CN** mantiene métricas relativamente superiores en comparación con las demás categorías, aunque degradadas respecto a las observadas en la validación. El modelo clasifica correctamente 46 de los 59 casos CN, igualando en términos absolutos al modelo de *transfer learning*; no obstante, este resultado se acompaña de una menor precisión, atribuible a una mayor tasa de falsos positivos.

Finalmente, la clase **MCI** presenta métricas consistentemente deficientes. El *recall* de 0.4667 indica que menos de la mitad de los casos son identificados correctamente, mientras que, en términos absolutos, el modelo clasifica adecuadamente 21 de los 45 casos MCI, lo que representa diez aciertos menos que ResNet3D. En conjunto, estos resultados confirman que la limitada capacidad representacional de VoxCNN compromete gravemente su desempeño en escenarios de generalización, especialmente en las categorías clínicamente más relevantes.

Promedios agregados en prueba:

- Promedios Macro: Precisión = 0.4433, Recall = 0.4393, F1-Score = 0.4337
- Promedios Micro: Precisión = Recall = F1-Score = 0.5763

Los promedios macro en prueba son dramáticamente inferiores a cualquier otro modelo evaluado, con valores que se aproximan a los de un clasificador aleatorio mejorado marginalmente. La diferencia de más de 22 puntos porcentuales en F1-score macro respecto a ResNet3D (0.4337 vs 0.6907) confirma la inadecuación de la arquitectura VoxCNN para la tarea de clasificación abordada.

5. Análisis de la matriz de confusión en el conjunto de prueba: La matriz de confusión en el conjunto de prueba (Figura 7.26) revela la magnitud extrema de las deficiencias del modelo en datos completamente independientes.

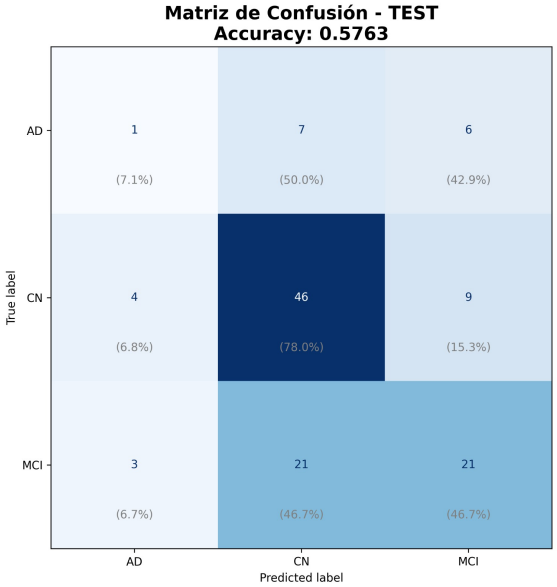


Figura 7.26: Matriz de confusión del modelo VoxCNN 3D en el conjunto de prueba.

Clasificaciones correctas:

- 1 de 14 casos AD (7.1 %) — reducción de −7 casos respecto a ResNet3D
- 46 de 59 casos CN (78.0 %) — mejora de +2 casos respecto a ResNet3D
- 21 de 45 casos MCI (46.7 %) — reducción de −10 casos respecto a ResNet3D

La diagonal de la matriz confirma el colapso del modelo para las clases minoritarias. La identificación de solo 1 caso AD de 14 representa un fallo sistémico de la arquitectura, mientras que la tasa de acierto en MCI (46.7 %) apenas supera el azar en un problema triclase.

6. Análisis de las curvas ROC: Las curvas ROC proporcionan una perspectiva adicional sobre las limitaciones discriminativas del modelo VoxCNN.

Curvas ROC en el conjunto de validación:

La Figura 7.27 presenta las curvas ROC para cada clase en validación.

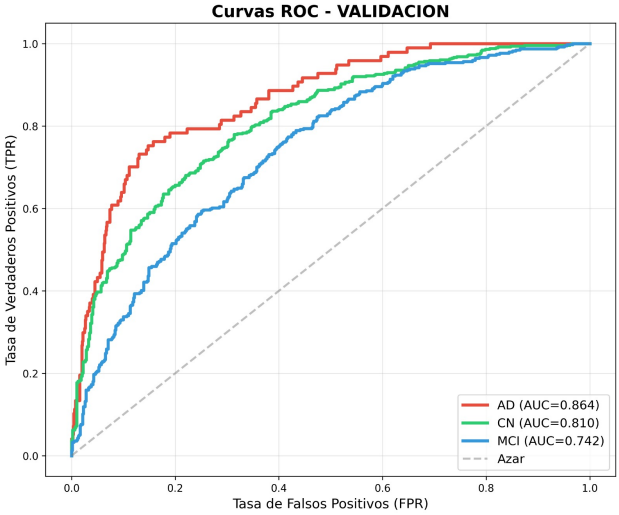


Figura 7.27: Curvas ROC del modelo VoxCNN 3D en el conjunto de validación.

El análisis comparativo con arquitecturas previas revela:

- **AD (AUC=0.864 vs 0.890 de ResNet3D):** Reducción de -0.026, indicando menor capacidad discriminativa. La curva roja muestra una elevación menos pronunciada en regiones de baja tasa de falsos positivos, sugiriendo dificultad para configurar el modelo en régimenes de alta especificidad manteniendo sensibilidad aceptable.

- **CN (AUC=0.810 vs 0.845 de ResNet3D):** Reducción de -0.035, con la curva verde mostrando un ascenso más gradual y menos consistente que en modelos previos.

- **MCI (AUC=0.742 vs 0.813 de ResNet3D):** Reducción de -0.071, ubicando la capacidad discriminativa en el límite inferior del rango aceptable. La curva azul se aproxima peligrosamente a la línea del clasificador aleatorio en varias regiones del espacio de umbrales.

Curvas ROC en el conjunto de prueba:

La Figura 7.28 muestra las curvas ROC en el conjunto de prueba, donde las limitaciones del modelo se manifiestan de manera aún más pronunciada.

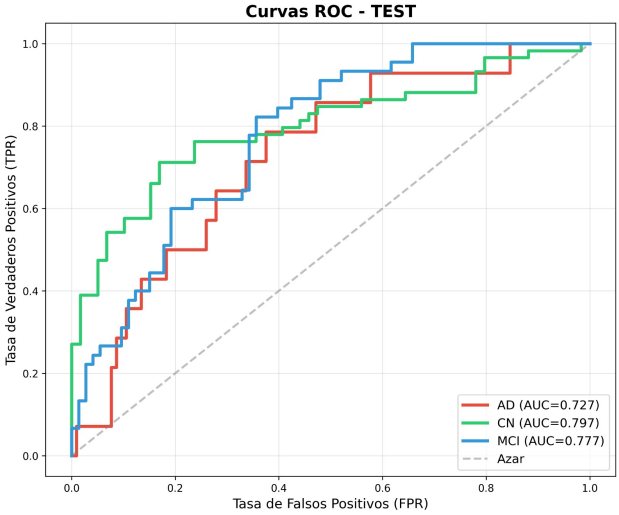


Figura 7.28: Curvas ROC del modelo VoxCNN 3D en el conjunto de prueba.

Los resultados en el conjunto de prueba muestran degradaciones alarmantes:

- **AD (AUC=0.727 vs 0.958 de ResNet3D):** Reducción catastrófica de −0.231, la mayor diferencia observada entre cualquier par de modelos para cualquier clase. El AUC de 0.727 apenas supera el umbral de utilidad clínica (0.70), indicando capacidad discriminativa marginal.

- **CN (AUC=0.797 vs 0.822 de ResNet3D):** Reducción de −0.025, la menos severa de las tres clases, consistente con el patrón de sesgo hacia CN observado en las matrices de confusión.

- **MCI (AUC=0.777 vs 0.798 de ResNet3D):** Reducción de −0.021, aunque partiendo de un valor ya bajo en validación.

La degradación de 13.7 puntos porcentuales en AUC para AD entre validación (0.864) y prueba (0.727) es particularmente preocupante, sugiriendo que el modelo no ha aprendido características robustas y generalizables para esta clase crítica.

7.3.2. Evaluación de los modelos de Machine Learning (ML)

7.3.2.1. Evaluación del modelo K-Nearest Neighbors (KNN)

1. Evaluación de la selección del hiperparámetro k: Se evaluó un rango de valores de k desde 1 hasta 30 mediante validación cruzada de 5 pliegues sobre el conjunto de entrenamiento. La Figura 7.29 presenta la evolución de la exactitud en función del número de vecinos considerados.

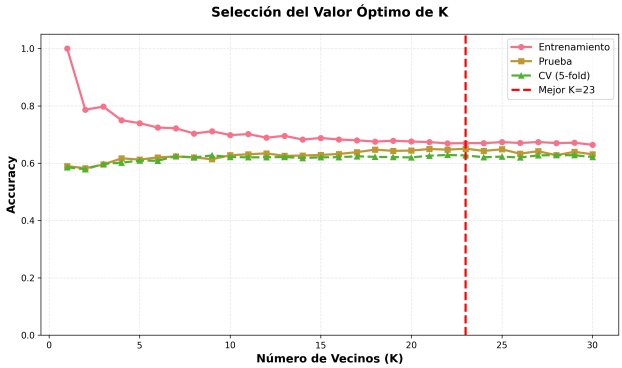


Figura 7.29: Selección del valor óptimo de k para el modelo KNN.

El análisis de la curva de optimización revela:

- Con $k = 1$, el modelo alcanza una exactitud de entrenamiento del 100 %, reflejando el comportamiento trivial en el que cada instancia se clasifica idénticamente a sí misma. Este valor disminuye rápidamente a medida que k aumenta, evidenciando la reducción del sobreajuste.

- Las tres curvas (entrenamiento, prueba y validación cruzada) convergen progresivamente a medida que k aumenta, estabilizándose en un rango entre 60–67 % para valores de k superiores a 15.

- El máximo desempeño en validación cruzada se alcanza con $k = 23$, donde la exactitud es del 62.71 %. Este valor representa un balance adecuado entre la capacidad de capturar la estructura local en los datos y la robustez frente al ruido.

- Para $k = 23$, la diferencia entre entrenamiento (66.99 %) y prueba (65.04 %) es de apenas 1.95 %, indicando una generalización apropiada sin sobreajuste significativo.

2. Desempeño global del modelo: El modelo KNN con $k = 23$ alcanzó una **exactitud global del 65.04 %** en el conjunto de prueba. La Figura 7.30 presenta una comparación de la exactitud entre los diferentes conjuntos de evaluación.

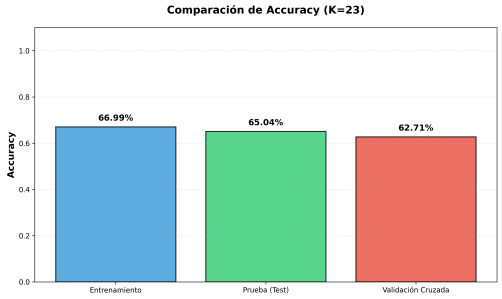


Figura 7.30: Comparación de exactitud del modelo KNN ($k = 23$) entre diferentes conjuntos de evaluación.

Métricas globales:

Cuadro 7.7: Métricas globales del modelo

Métrica	**Valor**
Exactitud en entrenamiento	66.99 %
Exactitud en prueba (test)	65.04 %
Exactitud en validación cruzada (5-fold)	62.71 %
Precisión macro	68.28 %
Recall macro	59.54 %
F1-score macro	62.10 %
F1-score ponderado	64.64 %
AUC promedio	0.8071

La consistencia entre las métricas de entrenamiento, prueba y validación cruzada, con diferencias menores al 5 %, indica que el modelo no sufrió sobreajuste y mantiene una capacidad de generalización razonable. El AUC promedio de 0.8071 sugiere una capacidad discriminativa buena.

3. Desempeño por clase La Figura 7.31 presenta una comparación visual de las métricas de precisión, recall y F1-score para cada categoría diagnóstica.

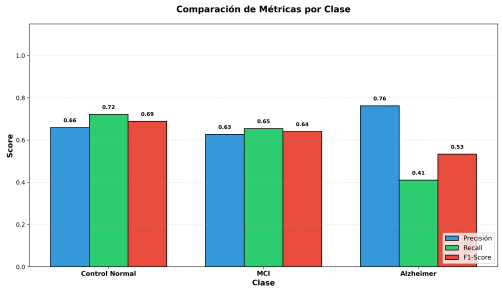


Figura 7.31: Comparación de métricas por clase para el modelo KNN.

Cuadro 7.8: Desempeño por clase

Métrica	**CN**	**MCI**	**AD**
Muestras en prueba	370	437	117
Precisión	65.93 %	62.72 %	76.19 %
Recall (Sensibilidad)	72.16 %	65.45 %	41.03 %
F1-Score	68.90 %	64.05 %	53.33 %
Especificidad	75.09 %	65.09 %	98.14 %
AUC	0.8031	0.6867	0.9317

El análisis del desempeño por clase evidencia comportamientos diferenciados entre las categorías diagnósticas. La clase **CN** presenta el mejor balance entre precisión y sensibilidad, con un *recall* de 72.16 %, lo que indica que aproximadamente siete de cada diez controles normales son identificados correctamente. En concordancia, el F1-score de 68.90 % refleja un desempeño sólido y equilibrado, mientras que el valor de AUC de 0.8031 confirma una buena capacidad discriminativa para separar esta clase de las categorías asociadas al deterioro cognitivo.

Por su parte, la clase **MCI** exhibe métricas intermedias y relativamente equilibradas. El *recall* de 65.45 % sugiere que cerca de dos tercios de los casos MCI son identificados correctamente, y el F1-score de 64.05 % resulta comparable al observado en la clase CN. No obstante, el AUC de 0.6867 es considerablemente inferior, lo que indica una mayor dificultad del modelo para discriminar esta categoría transicional frente a las clases adyacentes. Este comportamiento es consistente con la heterogeneidad clínica inherente al MCI, previamente documentada.

En contraste, la clase **AD** presenta el patrón más desequilibrado entre las métricas evaluadas. La precisión elevada (76.19 %) y la especificidad sobresaliente (98.14 %) evidencian que, cuando el modelo asigna la categoría AD, acierta en más de tres de cada cuatro casos y genera muy pocos falsos positivos. Sin embargo, el *recall* reducido (41.03 %) revela que solo cuatro de cada diez casos de Alzheimer son detectados, dejando inadvertidos aproximadamente el 59 % de los casos reales. Este desbalance se refleja en un F1-score de 53.33 %, el más bajo entre las tres clases.

Adicionalmente, el AUC de 0.9317, el más alto de todas las categorías, sugiere una capacidad discriminativa excelente para separar AD del resto a nivel de probabilidades. No obstante, este resultado indica que el umbral de decisión empleado es excesivamente conservador, priorizando la minimización de falsos positivos en detrimento de la sensibilidad.

4. Análisis de la matriz de confusión: La matriz de confusión (Figura 7.32) proporciona una visualización detallada de los patrones de clasificación y los errores del modelo KNN.

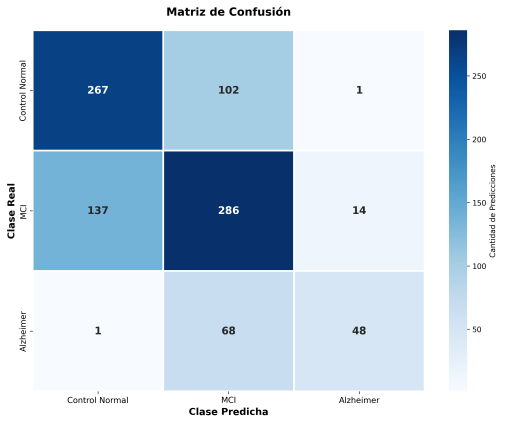


Figura 7.32: Matriz de confusión del modelo KNN en el conjunto de prueba.

Análisis diagonal (clasificaciones correctas):

- 267 de 370 casos CN correctamente clasificados (72.2 %)
- 286 de 437 casos MCI correctamente clasificados (65.5 %)
- 48 de 117 casos AD correctamente clasificados (41.0 %)

La diagonal confirma que CN presenta la mayor tasa de clasificación correcta, seguida por MCI, mientras que AD muestra el desempeño más deficiente con menos de la mitad de los casos identificados adecuadamente.

5. Análisis de las curvas ROC: Las curvas ROC proporcionan una caracterización completa del desempeño discriminativo del modelo KNN a través de diferentes umbrales de decisión. La Figura 7.33 presenta las curvas para cada clase.

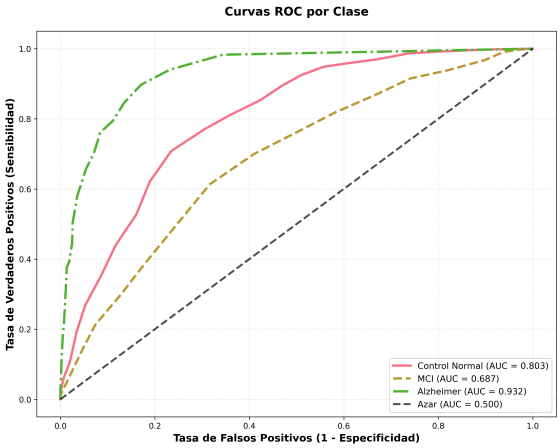


Figura 7.33: Curvas ROC del modelo KNN por clase.

El análisis de las curvas ROC revela:

- **Alzheimer (AUC=0.9317):** La curva verde muestra una elevación pronunciada y casi vertical en las regiones iniciales, alcanzando aproximadamente 75 % de sensibilidad con tasas de falsos positivos inferiores al 10 %. Este comportamiento excepcional confirma que el modelo posee una capacidad discriminativa excelente para separar AD de las otras categorías a nivel de puntuaciones de probabilidad. El AUC de 0.932 es superior incluso al de algunos modelos CNN para esta clase, lo que sugiere que las variables clínicas capturan información altamente predictiva del diagnóstico de Alzheimer. Sin embargo, la discrepancia entre este AUC elevado y el recall bajo (41 %) indica que el umbral de decisión por defecto (0.5) está mal calibrado para esta clase minoritaria.
- **Control Normal (AUC=0.8031):** La curva rosada presenta un ascenso sostenido y relativamente uniforme, alcanzando aproximadamente 72 % de sensibilidad (consistente con el recall reportado) con una tasa de falsos positivos cercana al 25 %. El AUC de 0.803 indica buena capacidad discriminativa, permitiendo configurar el modelo en diversos puntos de operación según las prioridades clínicas.
- **MCI (AUC=0.6867):** La curva amarilla muestra la menor separación respecto al clasificador aleatorio, con un ascenso más gradual y menos pronunciado que las otras clases. El AUC de 0.687 se aproxima al límite inferior del rango considerado aceptable (0.7), reflejando la dificultad inherente para discriminar esta categoría transicional de CN y AD utilizando las variables disponibles. Este comportamiento es consistente con la heterogeneidad clínica del MCI documentada en secciones anteriores.

7.3.2.2. Evaluación del modelo K-Nearest Neighbors con SMOTE

1. Evaluación de la aplicación del balanceo con SMOTE: La Figura 7.34 ilustra el impacto de SMOTE sobre la distribución de clases en el conjunto de entrenamiento.

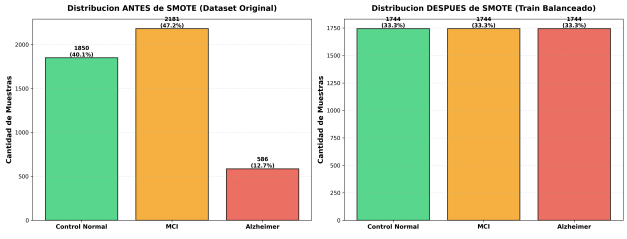


Figura 7.34: Comparación de la distribución de clases antes y después de aplicar SMOTE.

El balanceo alcanzado mediante SMOTE transformó un conjunto de entrenamiento original de 3,693 muestras en uno balanceado de 5,232 muestras, donde cada clase cuenta con exactamente 1,744 instancias. Este equilibrio elimina el sesgo inherente hacia la clase mayoritaria que afectó al modelo KNN estándar.

Optimización de hiperparámetros mediante Grid Search:

Se realizó una búsqueda de hiperparámetros, evaluando 72 combinaciones distintas de los siguientes parámetros:

Cuadro 7.9: Búsqueda y configuración óptima de hiperparámetros KNN

Hiperparámetro	**Valores explorados**	**Valor óptimo**
n_neighbors (k)	[1, 2, ..., 30]	5
weights	['uniform', 'distance']	'distance'
metric	['euclidean', 'manhattan']	'manhattan'
p	[1, 2]	1
Score de validación cruzada (5-fold): 75.44 %		

La Figura 7.35 muestra el impacto del parámetro k sobre el desempeño en el conjunto de prueba.

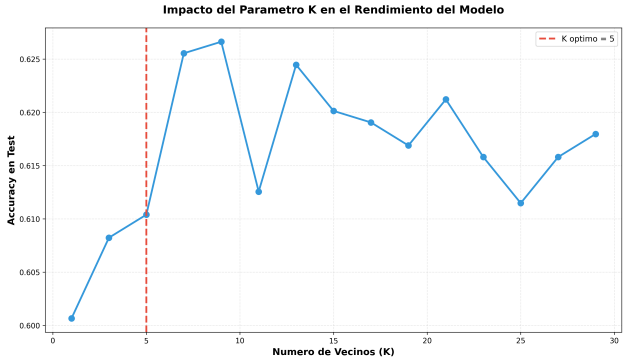


Figura 7.35: Impacto del parámetro k (número de vecinos) en la exactitud del modelo KNN con SMOTE sobre el conjunto de prueba.

Resulta notable que el valor óptimo $k = 5$ seleccionado mediante validación cruzada (75.44 %) corresponde a una región de desempeño relativamente bajo en el conjunto de prueba (aproximadamente 61 %). Este comportamiento sugiere una posible discrepancia entre los patrones aprendidos durante la validación cruzada sobre datos balanceados con SMOTE y el desempeño real sobre el conjunto de prueba no balanceado, un fenómeno que se discutirá posteriormente.

2. Desempeño global del modelo: El modelo KNN con SMOTE alcanzó una **exactitud global del 61.04 %** en el conjunto de prueba, lo que representa una **degradación de 4 puntos porcentuales respecto al modelo KNN estándar** (65.04 %). Este resultado contrarresta parcialmente la hipótesis inicial de que el balanceo de clases mejoraría el desempeño global del modelo.

La Figura 7.36 presenta una comparación de la exactitud entre diferentes conjuntos de evaluación, revelando patrones preocupantes de sobreajuste.

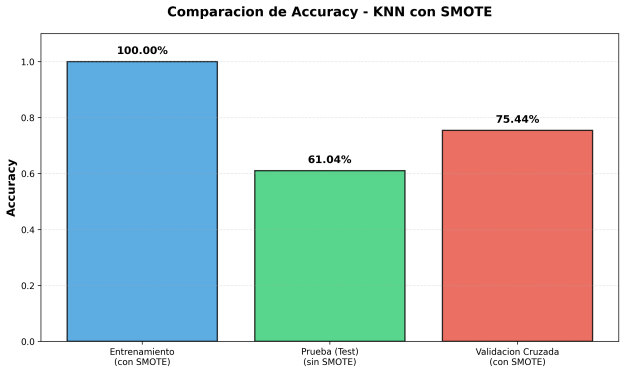


Figura 7.36: Comparación de exactitud del modelo KNN con SMOTE entre diferentes conjuntos de evaluación.

Métricas globales:

Cuadro 7.10: Métricas globales del modelo

Métrica	**Valor**
Exactitud en entrenamiento (con SMOTE)	100.00 %
Exactitud en prueba (sin SMOTE)	61.04 %
Exactitud en validación cruzada (con SMOTE)	75.44 % (±2.27 %)
Precisión macro	59.61 %
Recall macro	63.61 %
F1-score macro	61.07 %
F1-score ponderado	60.82 %
AUC promedio	0.7748

El análisis del desempeño del modelo entrenado con un conjunto balanceado mediante **SMOTE** evidencia un comportamiento marcadamente discordante entre las distintas etapas de evaluación. En particular, la exactitud perfecta (100 %) alcanzada en el conjunto de entrenamiento contrasta de forma abrupta con el desempeño obtenido en el conjunto de prueba (61.04 %), generando una brecha cercana a 39 puntos porcentuales. Este patrón constituye evidencia inequívoca de un sobreajuste severo, en el cual el modelo memoriza las características específicas de las muestras sintéticas generadas por SMOTE sin desarrollar representaciones robustas y generalizables a datos reales.

De manera paradójica, la validación cruzada realizada bajo el mismo esquema de balanceo arroja un *score* de 75.44 %, un valor significativamente superior al desempeño observado en el conjunto de prueba. Esta discrepancia se explica por el hecho de que la validación cruzada se llevó a cabo íntegramente sobre datos balanceados mediante SMOTE, donde todas las clases presentan una representación equivalente, mientras que el conjunto de prueba preserva el desbalanceo natural de la población. En consecuencia, este resultado pone de manifiesto una limitación metodológica crítica del uso de SMOTE, ya que la evaluación basada en validación cruzada puede conducir a estimaciones

excesivamente optimistas que no se traducen en un desempeño comparable sobre datos reales no balanceados.

3. Desempeño por clase: La Figura 7.37 presenta una comparación visual de las métricas de precisión, recall y F1-score para cada categoría diagnóstica, lo que permite evaluar el impacto específico de SMOTE sobre cada clase.

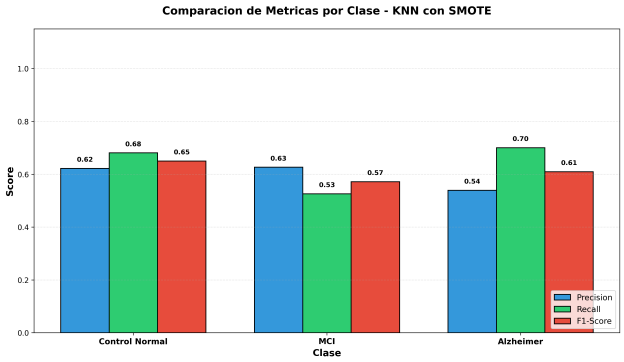


Figura 7.37: Comparación de métricas por clase para el modelo KNN con SMOTE.

Cuadro 7.11: Desempeño por clase

Métrica	**CN**	**MCI**	**AD**
Muestras en prueba	370	437	117
Precisión	62.22 % (−3.71 %)	62.67 % (−0.05 %)	53.95 % (−22.24 %)
Recall (Sensibilidad)	68.11 % (−4.05 %)	52.63 % (−12.82 %)	70.09 % (+29.06 %)
F1-Score	65.03 % (−3.87 %)	57.21 % (−6.84 %)	60.97 % (+7.64 %)
Especificidad	72.38 % (−2.71 %)	71.87 % (+6.78 %)	91.33 % (−6.81 %)
AUC	0.7763 (−0.0268)	0.6721 (−0.0146)	0.8759 (−0.0558)

Valores entre paréntesis: diferencia respecto a KNN estándar

El análisis del desempeño por clase tras la aplicación de **SMOTE** evidencia efectos diferenciados y, en algunos casos, contrapuestos entre las categorías diagnósticas. En la clase **CN**, se observan degradaciones moderadas pero consistentes en todas las métricas. En particular, el *recall* del 68.11 % indica que aproximadamente dos tercios de los controles normales son identificados correctamente, lo que representa una reducción cercana a cuatro puntos porcentuales respecto al modelo KNN estándar. Este comportamiento sugiere que el balanceo artificial, si bien favorece la detección de la clase minoritaria AD, lo hace parcialmente a expensas del desempeño en las clases originalmente mejor representadas.

La clase **MCI** resulta ser la más perjudicada por el balanceo forzado. El *recall* desciende del 65.45 % en el KNN estándar al 52.63 % con SMOTE, lo que corresponde a una reducción de casi trece puntos porcentuales. Dado que MCI constituye la clase mayoritaria en el conjunto original (47.2 %),

este resultado indica que el entrenamiento con representación equitativa de las clases conduce a fronteras de decisión que no reflejan adecuadamente su prevalencia real. En consecuencia, el F1-score de 57.21 % pone de manifiesto un desbalance entre una precisión relativamente conservada y una pérdida sustancial de sensibilidad.

En contraste, la clase **AD** exhibe el cambio más dramático y representa el objetivo principal de la aplicación de SMOTE. El *recall* experimenta un incremento notable de 29 puntos porcentuales, pasando de 41.03 % en el KNN estándar a 70.09 %, lo que implica que el modelo identifica correctamente aproximadamente siete de cada diez casos de Alzheimer (82 de 117), más que duplicando la tasa de detección previa. No obstante, esta ganancia en sensibilidad se produce a costa de un colapso en la precisión, que disminuye de 76.19 % a 53.95 %, evidenciando una elevada tasa de falsos positivos. De manera consistente, la especificidad también se reduce de 98.14 % a 91.33 %.

Adicionalmente, el AUC para la clase AD disminuye de 0.9317 a 0.8759, una reducción de 5.6 puntos porcentuales. Este resultado, aparentemente contradictorio frente al aumento del *recall*, sugiere que el uso de SMOTE desplaza el umbral de decisión hacia una mayor sensibilidad, pero degrada la capacidad discriminativa global del modelo a nivel de puntuaciones de probabilidad. Este efecto puede atribuirse a la distorsión de las fronteras de decisión inducida por las muestras sintéticas, las cuales no capturan plenamente la variabilidad intrínseca de los casos reales.

4. Análisis de la matriz de confusión: La matriz de confusión (Figura 7.38) proporciona una visualización detallada de cómo SMOTE redistribuyó los patrones de clasificación y error del modelo.

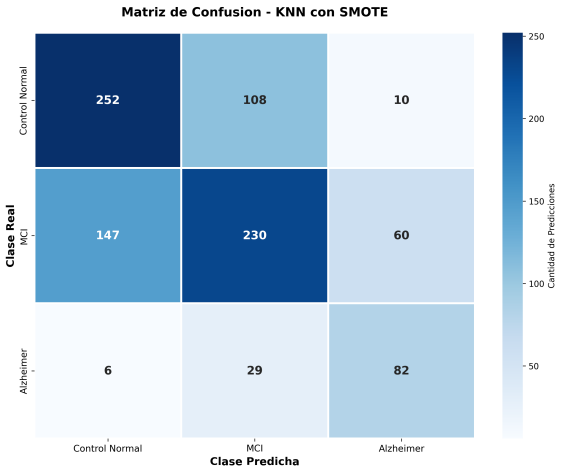


Figura 7.38: Matriz de confusión del modelo KNN con SMOTE en el conjunto de prueba.

Análisis diagonal (clasificaciones correctas):

- 252 de 370 casos CN correctamente clasificados (68.1 %) — reducción de −15 casos respecto al KNN estándar

- 230 de 437 casos MCI correctamente clasificados (52.6 %) — reducción de −56 casos respecto al KNN estándar
- 82 de 117 casos AD correctamente clasificados (70.1 %) — incremento de +34 casos respecto al KNN estándar

La diagonal confirma el trade-off fundamental introducido por SMOTE: mejora sustancial en la detección de AD (+34 casos, equivalente a +71 % de incremento relativo) a expensas de la degradación en CN (−15 casos, −5.6 %) y especialmente en MCI (−56 casos, −19.6 %).

5. Análisis de las curvas ROC: Las curvas ROC proporcionan información adicional sobre cómo SMOTE afectó la capacidad discriminativa del modelo a través de diferentes umbrales de decisión. La Figura 7.39 presenta las curvas para cada clase.

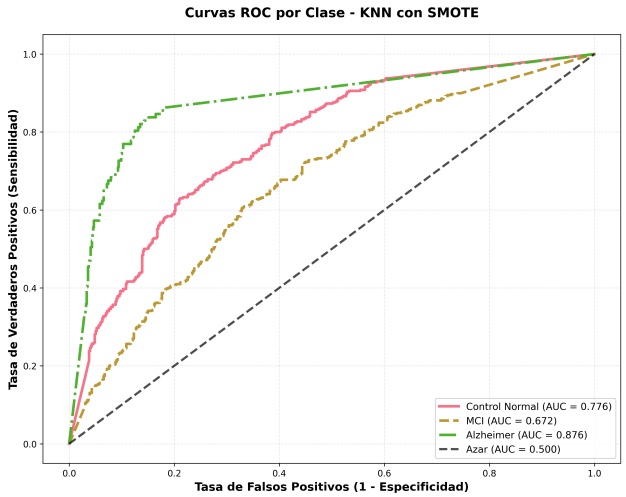


Figura 7.39: Curvas ROC del modelo KNN con SMOTE por clase.

El análisis comparativo de las curvas ROC con KNN estándar revela lo siguiente:

- **Alzheimer (AUC=0.8759 vs 0.9317 en KNN estándar):** Reducción de 5.6 puntos porcentuales en AUC, confirmando que SMOTE, a pesar de mejorar el recall, degradó la capacidad discriminativa a nivel de puntuaciones de probabilidad. La curva verde muestra menor elevación en las regiones iniciales comparada con KNN estándar, indicando que el modelo requiere aceptar tasas más altas de falsos positivos para alcanzar sensibilidades equivalentes.
- **Control Normal (AUC=0.7763 vs 0.8031 en KNN estándar):** Reducción de 2.7 puntos porcentuales. La curva rosada muestra un ascenso ligeramente menos pronunciado, reflejando la degradación moderada en todas las métricas para esta clase.

- **MCI (AUC=0.6721 vs 0.6867 en KNN estándar):** Reducción de 1.5 puntos porcentuales, manteniendo al modelo en el límite inferior del rango aceptable. El MCI continúa siendo la clase más difícil de discriminar, independientemente de la aplicación de SMOTE.

La degradación consistente de los AUCs en las tres clases sugiere que SMOTE, si bien reequilibra las tasas de detección entre clases mediante la recalibración de umbrales de decisión, lo hace a expensas de la calidad general de las puntuaciones de probabilidad y la separabilidad entre categorías.

7.3.2.3. Evaluación del modelo Naive Bayes

1. Desempeño global del modelo: El modelo Naive Bayes alcanzó una **exactitud global del 60.39 %** en el conjunto de prueba, representando el **desempeño más bajo entre todos los modelos de ML evaluados hasta este punto**, con una degradación de 4.65 puntos porcentuales respecto al KNN estándar (65.04 %) y de 0.65 puntos respecto al KNN con SMOTE (61.04 %).

La Figura 7.40 presenta una comparación de la exactitud entre diferentes conjuntos de evaluación, revelando un patrón excepcional de consistencia.

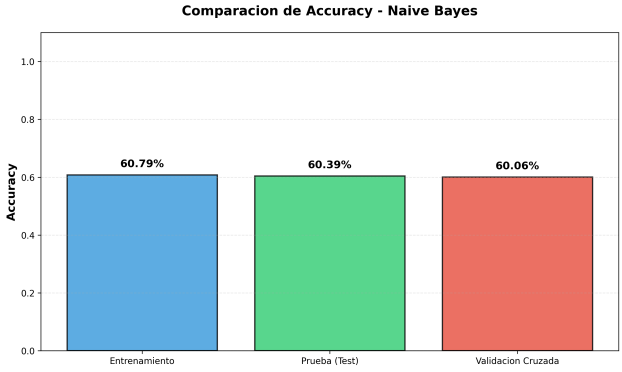


Figura 7.40: Comparación de exactitud del modelo Naive Bayes entre diferentes conjuntos de evaluación.

Métricas globales:

Cuadro 7.12: Métricas globales del modelo

Métrica	Valor
Exactitud en entrenamiento	60.79 %
Exactitud en prueba (test)	60.39 %
Exactitud en validación cruzada	60.06 % (±1.63 %)
Precisión macro	58.93 %
Recall macro	58.45 %
F1-score macro	58.27 %
F1-score ponderado	59.91 %
AUC promedio	0.7699

El análisis del comportamiento del modelo evidencia una generalización consistente entre el conjunto de entrenamiento y los datos no vistos, sin indicios de memorización de idiosincrasias específicas del conjunto de entrenamiento. La consistencia prácticamente perfecta entre los distintos conjuntos de evaluación sugiere que el modelo ha alcanzado el límite de su capacidad de representación bajo las asunciones de independencia condicional que lo caracterizan, sin margen adicional para capturar mayor complejidad presente en los datos.

Esta estabilidad resulta particularmente relevante considerando que la asunción de independencia condicional se encuentra claramente violada en el contexto del problema abordado, dado que las variables clínicas empleadas presentan correlaciones conocidas. No obstante, las distribuciones estimadas por el modelo logran capturar una estructura suficiente para generalizar de manera consistente, aunque a un nivel de desempeño absoluto inferior en comparación con modelos de mayor capacidad representacional.

Adicionalmente, la desviación estándar obtenida en la validación cruzada (1.63 %) es la más baja entre todos los modelos evaluados, lo que confirma la elevada robustez y estabilidad del clasificador probabilístico. En conjunto, estos resultados indican que el modelo prioriza la consistencia y la fiabilidad en la generalización sobre la maximización del desempeño, constituyéndose como un referente estable, aunque limitado, dentro del conjunto de enfoques analizados.

2. Desempeño por clase: La Figura 7.41 presenta una comparación visual de las métricas de precisión, recall y F1-score para cada categoría diagnóstica.

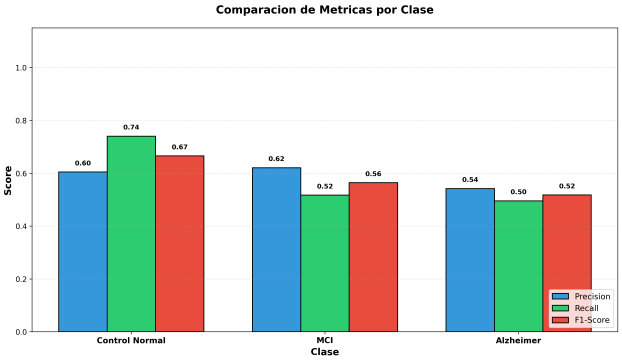


Figura 7.41: Comparación de métricas por clase para el modelo Naive Bayes.

Cuadro 7.13: Desempeño por clase

Métrica	**CN**	**MCI**	**AD**
Muestras en prueba	370	437	117
Precisión	60.49 % (−5.44 %)	62.09 % (−0.63 %)	54.21 % (−21.98 %)
Recall (Sensibilidad)	74.05 % (+1.89 %)	51.72 % (−13.73 %)	49.57 % (+8.54 %)
F1-Score	66.59 % (−2.31 %)	56.43 % (−7.62 %)	51.79 % (−1.54 %)
Especificidad	67.69 % (−7.40 %)	71.66 % (+6.57 %)	93.93 % (−4.21 %)
AUC	0.7653 (−0.0378)	0.6404 (−0.0463)	0.9040 (−0.0277)

Valores entre paréntesis: diferencia respecto a KNN estándar

El análisis del desempeño por clase del clasificador **Naive Bayes** revela un comportamiento heterogéneo entre las categorías diagnósticas, condicionado por las asunciones probabilísticas del modelo. La clase **CN** exhibe el mejor desempeño relativo, con un *recall* de 74.05 %, ligeramente superior al obtenido por KNN estándar (72.16 %), lo que indica que aproximadamente tres cuartas partes de los controles normales son identificados correctamente. No obstante, la precisión reducida (60.49 %) evidencia una tasa elevada de falsos positivos, en la cual sujetos MCI o AD son clasificados erróneamente como CN. En consecuencia, el F1-score de 66.59 %, aunque el más alto dentro de este modelo, resulta inferior al de KNN estándar (68.90 %) y significativamente menor al alcanzado por los modelos CNN.

Por su parte, la clase **MCI** presenta el desempeño más limitado bajo el esquema de Naive Bayes. El *recall* del 51.72 % representa una degradación cercana a 14 puntos porcentuales respecto al KNN estándar, indicando que menos de la mitad de los casos de MCI son detectados correctamente. En términos absolutos, 211 de los 437 casos (48.3 %) son clasificados erróneamente como CN o AD. Asimismo, el AUC de 0.6404 se sitúa por debajo del umbral generalmente considerado útil en contextos clínicos (0.70), lo que sugiere una capacidad discriminativa marginal para esta categoría.

En la clase **AD**, Naive Bayes muestra métricas moderadas y relativamente balanceadas. El *recall* de 49.57 % supone una mejora de aproximadamente 8.5 puntos porcentuales frente a KNN estándar (41.03 %); sin embargo, este valor permanece en un nivel clínicamente insuficiente, dado que solo se detecta correctamente cerca de la mitad de los casos de Alzheimer (58 de 117). Esta sensibilidad es sustancialmente inferior a la alcanzada por KNN con SMOTE (70.09 %), lo que indica que Naive Bayes no logra una detección efectiva de la clase minoritaria sin estrategias explícitas de balanceo.

Adicionalmente, la precisión de 54.21 % resulta notablemente baja, con una caída superior a 22 puntos porcentuales respecto a KNN estándar (76.19 %), lo que implica que casi la mitad de las predicciones de AD corresponden a falsos positivos. Este patrón contrasta con el comportamiento del KNN estándar, caracterizado por alta precisión y baja sensibilidad, mientras que Naive Bayes presenta ambas métricas en niveles moderadamente bajos.

Asimismo, el AUC de 0.9040 para la clase AD constituye el segundo valor más alto entre todos los modelos de aprendizaje automático evaluados, solo superado por el KNN estándar (0.9317). Este resultado sugiere que las probabilidades estimadas por Naive Bayes poseen una capacidad discriminativa elevada a nivel de puntuaciones continuas; sin embargo, el umbral de decisión por defecto (0.5) no resulta óptimo para maximizar el equilibrio entre precisión y sensibilidad.

3. Análisis de la matriz de confusión: La matriz de confusión (Figura 7.42) proporciona una visualización detallada de los patrones de clasificación y error del modelo Naive Bayes.

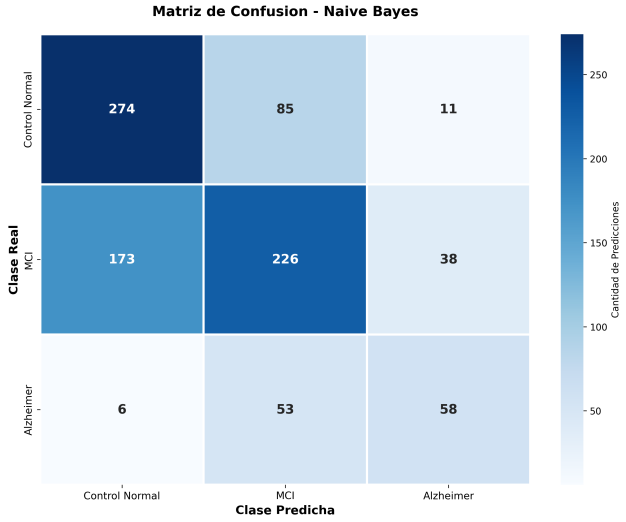


Figura 7.42: Matriz de confusión del modelo Naive Bayes en el conjunto de prueba.

Análisis diagonal (clasificaciones correctas):

- 274 de 370 casos CN correctamente clasificados (74.1 %) — incremento de +7 casos respecto a KNN estándar
- 226 de 437 casos MCI correctamente clasificados (51.7 %) — reducción de −60 casos respecto a KNN estándar
- 58 de 117 casos AD correctamente clasificados (49.6 %) — incremento de +10 casos respecto a KNN estándar

La diagonal revela un patrón mixto: mejora moderada en CN (+7 casos) y AD (+10 casos) respecto al KNN estándar, pero una degradación severa en MCI (−60 casos, equivalente a −21 % de reducción relativa). Este comportamiento indica que Naive Bayes desarrolló fronteras de decisión que favorecen ligeramente las clases extremas (CN y AD) a expensas de la categoría intermedia MCI.

4. Análisis de las curvas ROC: Las curvas ROC proporcionan una caracterización completa de la capacidad discriminativa del modelo a través de diferentes umbrales de decisión. La Figura 7.43 presenta las curvas para cada clase.

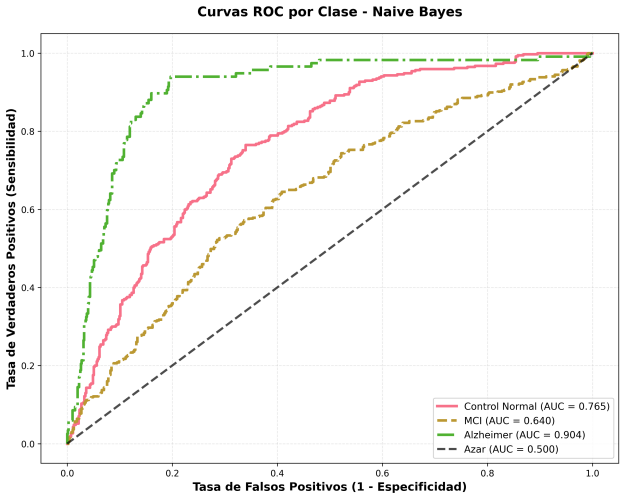


Figura 7.43: Curvas ROC del modelo Naive Bayes por clase.

El análisis de las curvas ROC revela:

- **Alzheimer (AUC=0.9040):** La curva verde muestra una elevación pronunciada y casi vertical en las regiones iniciales, alcanzando aproximadamente del 70 al 80 % de sensibilidad, con tasas de falsos positivos inferiores al 10 %. Este AUC, el segundo más alto entre todos los modelos de ML, confirma que Naive Bayes posee una capacidad discriminativa sobresaliente para AD a nivel de puntuaciones probabilísticas, aunque el umbral de decisión por defecto no optimiza el balance entre precisión y sensibilidad. El AUC es ligeramente inferior al de KNN estándar (0.9317) pero superior al de KNN con SMOTE (0.8759).
- **Control Normal (AUC=0.7653):** La curva rosada presenta un ascenso sostenido y relativamente uniforme, reflejando una capacidad discriminativa buena pero no excepcional. El AUC es inferior tanto al de KNN estándar (0.8031) como al de KNN con SMOTE (0.7763), lo que sugiere que las suposiciones de independencia condicional limitan la capacidad del modelo para capturar completamente las relaciones multivariadas que permiten una discriminación óptima de CN.
- **MCI (AUC=0.6404):** La curva amarilla muestra el ascenso más gradual y una menor separación respecto al clasificador aleatorio entre todas las clases. El AUC de 0.640 se encuentra por debajo del umbral de utilidad clínica (0.70), representando el valor más bajo observado entre todos los modelos de ML para esta categoría. Este resultado confirma que MCI constituye un desafío fundamental para Naive Bayes, donde las distribuciones probabilísticas estimadas bajo independencia condicional no logran capturar la complejidad de esta categoría transicional.

La discrepancia entre el AUC elevado para AD (0.904) y el recall moderado (49.6 %) confirma que el umbral de decisión por defecto (0.5) no está optimizado para esta clase. Una reconfiguración del umbral, reduciendo el valor requerido para clasificar como AD, podría incrementar sustancialmente

el recall, aceptando una disminución moderada en la precisión, similar a la estrategia implícitamente aplicada por KNN con SMOTE.

7.3.2.4. Evaluación del modelo Random Forest con SMOTE

1. Evaluación del balance con SMOTE: La Figura 7.44 ilustra el impacto de SMOTE sobre la distribución de clases.

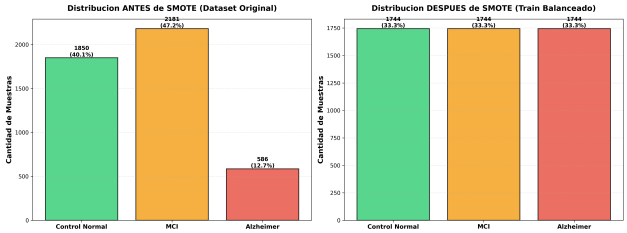


Figura 7.44: Comparación de la distribución de clases antes y después de aplicar SMOTE para Random Forest.

Optimización exhaustiva de hiperparámetros mediante Grid Search:

Se realizó una búsqueda evaluando 576 combinaciones distintas de los siguientes hiperparámetros mediante validación cruzada de 5 pliegues sobre el conjunto de entrenamiento balanceado con SMOTE:

Cuadro 7.14: Búsqueda y configuración óptima de hiperparámetros

Hiperparámetro	**Valores explorados**	**Valor óptimo**
n_estimators	[50, 100, 150, 200, 250, 300]	100
max_depth	[None, 10, 20, 30]	None
min_samples_split	[2, 5, 10]	2
min_samples_leaf	[1, 2, 4]	1
max_features	['sqrt', 'log2']	'sqrt'
bootstrap	[True, False]	False

Score de validación cruzada: 77.05 % (±3.61 %)

La Figura 7.45 muestra el impacto del número de árboles sobre el desempeño en el conjunto de prueba.

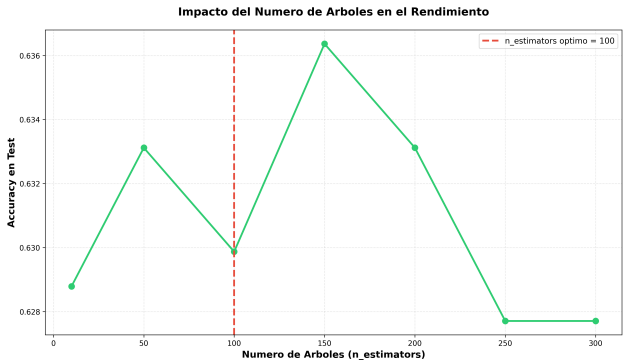


Figura 7.45: Impacto del número de árboles (n_estimators) en la exactitud del modelo Random Forest sobre el conjunto de prueba.

Resulta notable que el valor óptimo seleccionado mediante validación cruzada de 100 árboles no corresponde al máximo absoluto observado en el conjunto de prueba de aproximadamente 150 árboles con 63.7 %. Esta discrepancia refleja la variabilidad inherente entre conjuntos y sugiere que diferencias de desempeño menores al 1 % entre configuraciones son probablemente atribuibles a ruido estadístico más que a diferencias sistemáticas en la capacidad predictiva.

2. Desempeño global del modelo: La Figura 7.46 presenta una comparación de la exactitud entre diferentes conjuntos de evaluación, revelando un patrón similar al observado en KNN con SMOTE.

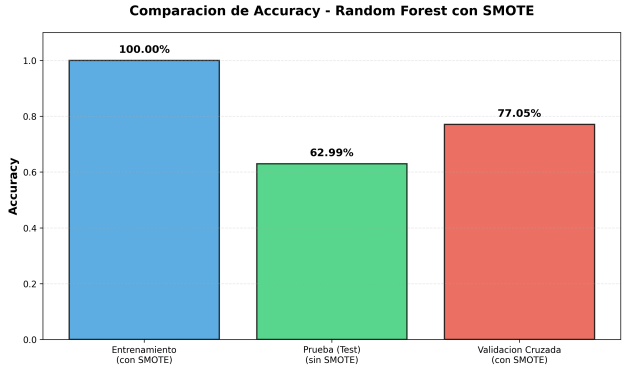


Figura 7.46: Comparación de exactitud del modelo Random Forest con SMOTE entre diferentes conjuntos de evaluación.

Cuadro 7.15: Métricas globales del modelo

Métrica	Valor
Exactitud en entrenamiento (con SMOTE)	100.00 %
Exactitud en prueba (sin SMOTE)	62.99 %
Exactitud en validación cruzada (con SMOTE)	77.05 % (±3.61 %)
Precisión macro	62.88 %
Recall macro	63.55 %
F1-score macro	63.20 %
F1-score ponderado	62.98 %
AUC promedio	0.8068

El patrón de resultados obtenido es consistente con el comportamiento observado previamente en KNN con SMOTE, evidenciando un claro compromiso entre la capacidad de ajuste y la generalización. La exactitud perfecta alcanzada en el conjunto de entrenamiento (100 %) indica que el ensamble de 100 árboles, configurado sin restricciones de profundidad, memorizó completamente el conjunto de datos balanceado mediante SMOTE. Este fenómeno es característico de Random Forests con configuraciones altamente permisivas (*max_ depth=None*, *min_ samples_ leaf=1*), en las que los árboles pueden crecer hasta profundidades elevadas, capturando incluso variaciones específicas introducidas por las muestras sintéticas.

Sin embargo, este elevado desempeño en entrenamiento contrasta con una brecha pronunciada entre entrenamiento y prueba, cuantificada en 37.01 puntos porcentuales, lo que confirma la presencia de un sobreajuste severo. El modelo parece haber aprendido patrones idiosincráticos asociados a las muestras generadas por SMOTE que no generalizan adecuadamente cuando se evalúa sobre datos reales con distribución desbalanceada. Esta falta de generalización se ve reforzada por la discrepancia sustancial entre el desempeño en validación cruzada y el conjunto de prueba: mientras que la validación cruzada sobre datos balanceados alcanza una exactitud del 77.05 %, el desempeño en prueba desciende a 62.99 %. Esta diferencia de 14.06 puntos porcentuales reproduce el patrón observado en KNN con SMOTE y confirma que las evaluaciones realizadas sobre conjuntos sintéticamente balanceados tienden a producir estimaciones optimistas que no se traducen en un rendimiento equivalente bajo la distribución original de los datos.

En términos de estabilidad, la desviación estándar del 3.61 % obtenida en la validación cruzada puede considerarse moderada. Este valor es superior al observado en Naive Bayes (1.63 %), pero inferior al reportado para KNN con SMOTE en algunas configuraciones, lo que sugiere que el ensamble mantiene una estabilidad razonable a través de los distintos pliegues del conjunto balanceado, aunque dicha estabilidad no es suficiente para mitigar los efectos del sobreajuste inducido por el balanceo sintético.

3. Desempeño por clase: La Figura 7.47 presenta una comparación visual de las métricas de precisión, recall y F1-score para cada categoría diagnóstica.

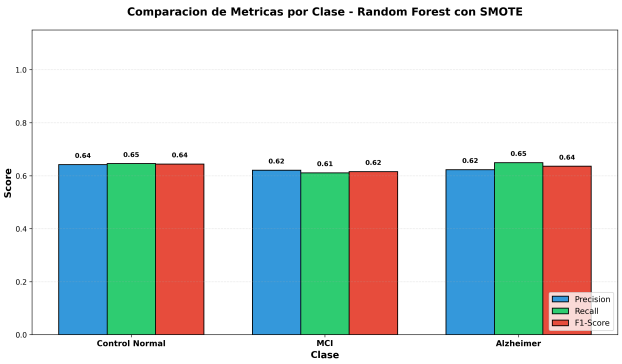


Figura 7.47: Comparación de métricas por clase para el modelo Random Forest con SMOTE.

Cuadro 7.16: Desempeño por clase

Métrica	**CN**	**MCI**	**AD**
Muestras en prueba	370	437	117
Precisión	64.25 % (−1.68 %)	62.09 % (−0.63 %)	62.30 % (−13.89 %)
Recall (Sensibilidad)	64.59 % (−7.57 %)	61.10 % (−4.35 %)	64.96 % (+23.93 %)
F1-Score	64.42 % (−4.48 %)	61.59 % (−2.46 %)	63.60 % (+10.27 %)
Especificidad	75.99 % (+0.90 %)	66.53 % (+1.44 %)	94.30 % (−3.84 %)
AUC	0.7951 (−0.0080)	0.6943 (+0.0076)	0.9309 (+0.0008)

Valores entre paréntesis: diferencia respecto a KNN estándar

La clase CN presenta un desempeño balanceado, con valores de precisión y *recall* prácticamente equivalentes (64.25 % y 64.59 %), lo que evidencia un equilibrio entre falsos positivos y falsos negativos. El modelo identifica correctamente el 64.6 % de los controles normales, un rendimiento inferior al de KNN estándar, pero superior al obtenido por KNN con SMOTE. El F1-score de 64.42 % confirma este comportamiento equilibrado, aunque refleja una ligera degradación respecto a los modelos de mejor desempeño global.

En la categoría MCI, Random Forest exhibe el rendimiento más bajo entre las tres clases, con métricas concentradas alrededor del 61 %. No obstante, el *recall* de 61.10 % representa una mejora sustancial frente a KNN con SMOTE y Naive Bayes, aunque permanece por debajo del desempeño alcanzado por KNN estándar. Destaca que el AUC de 0.6943 es el más alto observado para MCI entre todos los modelos de *machine learning* evaluados, lo que sugiere una mejor capacidad para capturar la estructura discriminativa de esta clase transicional a nivel de puntuaciones de probabilidad.

Por su parte, la clase AD muestra el patrón de mejora más relevante, con un *recall* del 64.96 % que supera ampliamente al de KNN estándar y se aproxima al obtenido por KNN con SMOTE. En términos absolutos, el modelo identifica correctamente 76 de los 117 casos de Alzheimer, evidenciando un incremento significativo en sensibilidad sin recurrir a balanceo sintético. Aunque la precisión disminuye respecto a KNN estándar, el valor alcanzado (62.30 %) es superior al observado en KNN con SMOTE y Naive Bayes, indicando un compromiso más favorable entre sensibilidad y

especificidad.

Finalmente, el AUC para la clase AD alcanza un valor de 0.9309, prácticamente idéntico al de KNN estándar y superior al de los modelos restantes, lo que confirma que Random Forest mantiene una capacidad discriminativa excelente para la detección del Alzheimer a nivel de puntuaciones de probabilidad, aun en presencia de un desbalance de clases.

4. Análisis de la matriz de confusión: La matriz de confusión (Figura 7.48) proporciona una visualización detallada de los patrones de clasificación y error del modelo Random Forest.

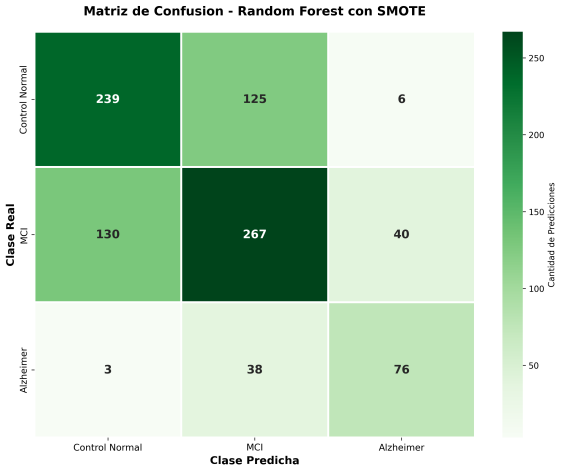


Figura 7.48: Matriz de confusión del modelo Random Forest con SMOTE en el conjunto de prueba.

Análisis diagonal (clasificaciones correctas):

- 239 de 370 casos CN correctamente clasificados (64.6 %) — reducción de −28 casos respecto a KNN estándar
- 267 de 437 casos MCI correctamente clasificados (61.1 %) — reducción de −19 casos respecto a KNN estándar
- 76 de 117 casos AD correctamente clasificados (65.0 %) — incremento de +28 casos respecto a KNN estándar

La diagonal revela el trade-off fundamental de SMOTE aplicado a Random Forest: mejora dramática en AD (+28 casos, +58 % de incremento relativo respecto a KNN estándar) a expensas de degradaciones moderadas en CN (−28 casos, −10.5 %) y MCI (−19 casos, −6.6 %). Resulta notable que el número de casos ganados en AD (+28) compensa exactamente el número de casos perdidos en CN (−28), sugiriendo una redistribución del desempeño más que una mejora neta global.

5. Análisis de la importancia de las características: Una ventaja distintiva de Random Forest es su capacidad para cuantificar la importancia relativa de cada característica en el proceso de clasificación. La Figura 7.49 presenta el ranking de las 11 características utilizadas.

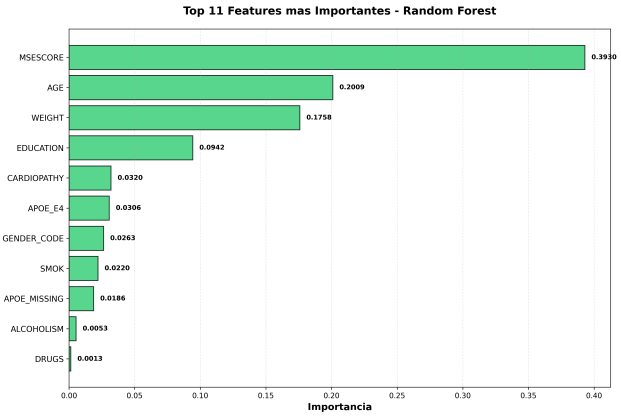


Figura 7.49: Importancia de características en el modelo Random Forest.

El análisis de importancia revela:
- Características dominantes:

Cuadro 7.17: Importancia de predictores en la clasificación de estadios cognitivos

Predictor	**Importancia**	**Interpretación clínica**
MMSE	39.3 %	Medida estandarizada del estado cognitivo global, diseñada específicamente para detectar deterioro cognitivo y demencia. Captura información directamente relacionada con el fenotipo clínico.
Edad	20.1 %	Refleja la naturaleza asociada al envejecimiento de la enfermedad de Alzheimer. El riesgo se incrementa exponencialmente después de los 65 años.
Peso	17.6 %	Podría reflejar asociaciones con estado de salud general, comorbilidades metabólicas o pérdida de peso característica de etapas avanzadas de demencia.
Educación	9.4 %	Predictor de reserva cognitiva. Mayor escolaridad se asocia con mayor resistencia a manifestaciones clínicas de patología neurodegenerativa.

La concentración del 77.6 % de importancia en las cuatro características numéricas (MMSE, edad, peso, educación) sugiere que estas variables capturan la mayoría de la información discriminativa disponible en el conjunto de datos.

- Características secundarias:

Las variables binarias presentan una contribución globalmente marginal al desempeño del modelo, con importancias individuales inferiores al 3.5 %. Entre ellas, la presencia del alelo APOE $\varepsilon 4$, reconocido como el factor de riesgo genético más establecido para la enfermedad de Alzheimer, alcanza una importancia relativa del 3.1 %, lo que indica un aporte modesto pero no despreciable al proceso de clasificación. En contraste, las variables asociadas a comorbilidades y hábitos de vida, incluyendo cardiopatía, tabaquismo, alcoholismo y consumo de drogas, exhiben contribuciones mínimas inferiores al 2.5 %, lo que sugiere que su capacidad predictiva es limitada o que su información se encuentra altamente correlacionada con otras variables de mayor relevancia incluidas en el modelo.

6. Análisis de las curvas ROC: Las curvas ROC proporcionan una caracterización completa de la capacidad discriminativa del modelo Random Forest. La Figura 7.50 presenta las curvas para cada clase.

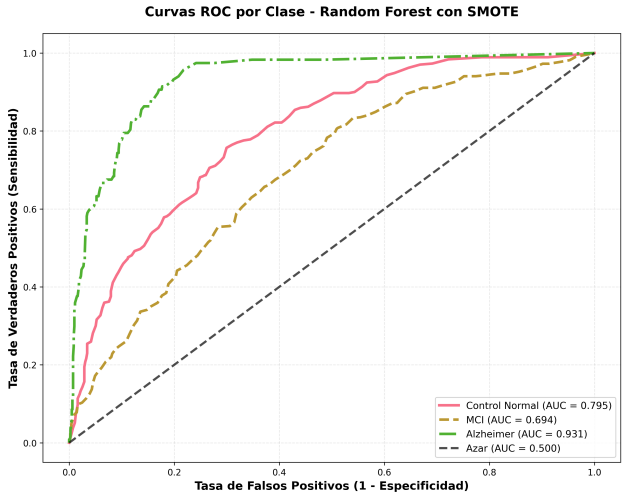


Figura 7.50: Curvas ROC del modelo Random Forest con SMOTE por clase.

El análisis comparativo de las curvas ROC con otros modelos de ML revela:

- **Alzheimer (AUC=0.9309):** La curva verde muestra una elevación pronunciada en regiones de baja tasa de falsos positivos, alcanzando aproximadamente 70–80 % de sensibilidad con menos del 10 % de falsos positivos. El AUC de 0.931 es prácticamente idéntico al KNN estándar (0.9317, −0.0008) y superior tanto al KNN con SMOTE (0.8759, +0.0550) como al Naive Bayes (0.9040, +0.0269), confirmando que Random Forest mantiene una capacidad discriminativa excelente para AD a nivel probabilístico, mientras mejora simultáneamente el recall mediante la recalibración de umbrales inducida por SMOTE.
- **Control Normal (AUC=0.7951):** La curva rosada muestra una buena separación respecto al clasificador aleatorio. El AUC es inferior al KNN estándar (0.8031, −0.0080) pero superior

al KNN con SMOTE (0.7763, +0.0188) y al Naive Bayes (0.7653, +0.0298), indicando una capacidad discriminativa intermedia que supera a modelos con SMOTE pero no alcanza el nivel del KNN sin balanceo.

- **MCI (AUC=0.6943):** La curva amarilla presenta el ascenso más gradual, apenas superando el umbral de utilidad clínica (0.70). El AUC es superior al KNN estándar (0.6867, +0.0076), al KNN con SMOTE (0.6721, +0.0222) y significativamente superior al Naive Bayes (0.6404, +0.0539), representando el mejor desempeño discriminativo para MCI entre todos los modelos de ML evaluados. Este resultado sugiere que la capacidad de Random Forest para capturar relaciones no lineales complejas proporciona ventajas para discriminar esta categoría heterogénea y transicional.

7.4. Evaluación del modelo híbrido implementado

1. Desempeño global del modelo: El modelo híbrido alcanzó una **exactitud global del 77.12 %** en el conjunto de prueba, representando el **mejor desempeño entre todos los enfoques evaluados en este estudio**. Este resultado evidencia una mejora sustancial respecto a los componentes individuales:

- **+10.13 puntos porcentuales** respecto al modelo Transfer Learning standalone (70.34 %)
- **+14.13 puntos porcentuales** respecto al Random Forest con variables clínicas (62.99 %)
- **+6.78 puntos porcentuales** respecto al mejor modelo CNN individual evaluado (TL ResNet-10)

Esta mejora no puede atribuirse a la simple agregación de información, sino que evidencia un **efecto sinérgico** donde la integración multimodal permite que el clasificador capture patrones diagnósticos más robustos que los disponibles en cada modalidad por separado.

Métricas globales:

Cuadro 7.18: Métricas globales del modelo híbrido

Métrica	Valor
Exactitud global	77.12 %
Precisión macro	84.92 %
Recall macro	65.77 %
Especificidad macro	85.50 %
F1-score macro	69.59 %

Los promedios macro, que otorgan igual peso a cada clase independientemente de su prevalencia, confirman que el modelo híbrido mantiene un desempeño equilibrado entre categorías, con una precisión macro superior al 84 % y una especificidad media del 85.5 %, indicando una excelente capacidad para minimizar falsos positivos a nivel global.

2. Desempeño por clase diagnóstica: La Tabla 7.19 presenta un análisis comparativo del desempeño del modelo híbrido para cada categoría diagnóstica, incluyendo comparaciones con los mejores modelos individuales evaluados.

Cuadro 7.19: Desempeño del modelo híbrido por clase diagnóstica

Métrica	**AD**	**CN**	**MCI**
Resultados del modelo híbrido			
Precisión	100.00 %	73.68 %	81.08 %
Recall (Sensibilidad)	35.71 %	94.92 %	66.67 %
Especificidad	100.00 %	66.10 %	90.41 %
F1-Score	52.63 %	82.96 %	73.17 %
Comparación con Transfer Learning ResNet-10			
Δ Precisión	+30.77 %	−3.00 %	+18.86 %
Δ Recall	−28.58 %	+16.95 %	+4.45 %
Δ F1-Score	−14.04 %	+5.65 %	+10.95 %
Comparación con Random Forest + SMOTE			
Δ Precisión	+37.70 %	+9.43 %	+18.78 %
Δ Recall	−29.25 %	+30.33 %	+5.57 %
Δ F1-Score	−8.37 %	+17.93 %	+9.57 %
Información adicional			
Casos totales	14	59	45
Casos correctos	5	56	30
Tasa de acierto	35.71 %	94.92 %	66.67 %

Análisis por clase:

El modelo híbrido exhibe comportamientos claramente diferenciados según la clase diagnóstica, reflejando el impacto de la integración entre información de neuroimagen PET y variables clínicas–neuropsicológicas. En el caso de la enfermedad de Alzheimer (AD), el clasificador adopta una estrategia marcadamente conservadora, alcanzando una precisión y especificidad perfectas del 100 %, sin generar falsos positivos en el conjunto de prueba. Este resultado implica que todas las predicciones de AD corresponden inequívocamente a casos reales, lo cual es particularmente relevante desde la perspectiva de confirmación diagnóstica. No obstante, este desempeño se logra a expensas de una sensibilidad limitada, con un recall del 35.71 %, identificando únicamente 5 de los 14 casos de AD. En consecuencia, una proporción significativa de pacientes con Alzheimer es clasificada como MCI o CN, lo que evidencia un desplazamiento deliberado del compromiso precisión-recall hacia la maximización de certeza diagnóstica. Este patrón contrasta con los modelos unimodales, que exhibieron mayores niveles de sensibilidad pero menor precisión, y sugiere que la fusión prioriza la eliminación absoluta de falsos positivos sobre la detección exhaustiva de casos.

Por su parte, la clase de controles cognitivamente normales (CN) presenta el mejor desempeño global del modelo híbrido. El recall alcanza un valor del 94.92 %, con la identificación correcta de 56 de los 59 sujetos normales, superando ampliamente a los modelos base y evidenciando que la

integración multimodal permite capturar de manera más robusta los patrones asociados a la cognición preservada. Aunque la precisión es más moderada, del 73.68 %, el balance entre sensibilidad y especificidad se refleja en un F1-score del 82.96 %, el más alto entre todas las clases evaluadas. Los errores de clasificación se concentran principalmente en casos de MCI etiquetados como CN, un resultado coherente con la naturaleza transicional y heterogénea del deterioro cognitivo leve. Este desempeño sugiere que la combinación de la ausencia de hipometabolismo patológico en PET con perfiles clínicos favorables, como puntuaciones elevadas en MMSE y la ausencia de quejas cognitivas relevantes, proporciona evidencia suficiente para una identificación confiable de la normalidad cognitiva.

Finalmente, la clasificación del deterioro cognitivo leve (MCI) continúa representando el mayor desafío para el modelo, en concordancia con su reconocida heterogeneidad clínica y biológica. Aun así, el enfoque híbrido alcanza métricas balanceadas, con una precisión del 81.08 % y un recall del 66.67 %, identificando correctamente 30 de los 45 casos de MCI. Estos resultados superan de forma consistente a los modelos unimodales tanto en sensibilidad como en precisión, reflejándose en un F1-score sustancialmente superior. La elevada especificidad (90.41 %) indica que el clasificador rara vez confunde MCI con las categorías extremas de CN o AD, lo que sugiere que ha aprendido representaciones distintivas de este estadio intermedio del espectro cognitivo. Los errores restantes corresponden mayoritariamente a casos de MCI clasificados como CN, un fenómeno esperable dado que algunos sujetos presentan perfiles cercanos a la normalidad, progresión clínica variable o mecanismos de reserva cognitiva que atenúan la expresión del deterioro en las pruebas neuropsicológicas.

3. Análisis de la matriz de confusión: La Figura 7.51 presenta la matriz de confusión del modelo híbrido en el conjunto de prueba.

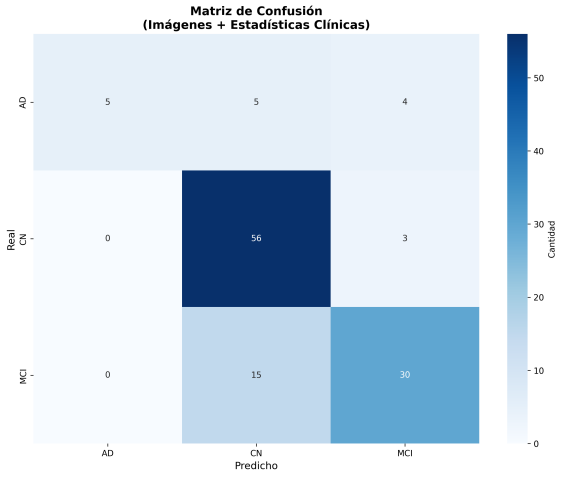


Figura 7.51: Matriz de confusión del modelo híbrido en el conjunto de prueba.

Análisis de la diagonal (clasificaciones correctas):

- 5 de 14 casos AD correctamente clasificados (35.71 %)

- 56 de 59 casos CN correctamente clasificados (94.92 %)
- 30 de 45 casos MCI correctamente clasificados (66.67 %)

4. Análisis de las curvas ROC: La Figura 7.52 presenta las curvas ROC para cada clase diagnóstica del modelo híbrido.

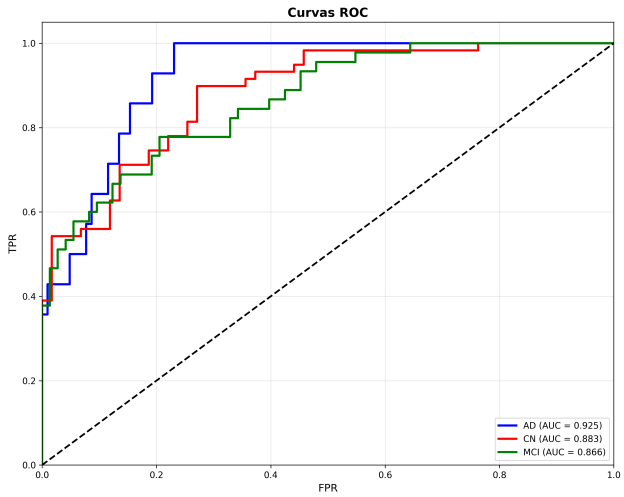


Figura 7.52: Curvas ROC del modelo híbrido por clase.

El análisis de las curvas ROC revela:

- **AD (AUC=0.925):** La curva azul muestra una elevación pronunciada y casi vertical en las regiones iniciales, alcanzando aproximadamente 50–60 % de sensibilidad con tasas de falsos positivos prácticamente nulas (<2 %). Este comportamiento confirma la capacidad discriminativa excepcional del modelo para AD a nivel de puntuaciones de probabilidad. El AUC de 0.925 supera al Transfer Learning (0.886) y se aproxima al Random Forest con SMOTE (0.931), indicando que la fusión multimodal mantiene la excelencia discriminativa mientras optimiza el balance precision-recall mediante criterios más estrictos.
- **CN (AUC=0.883):** La curva roja presenta un ascenso sostenido y consistente, reflejando la capacidad del modelo para identificar controles normales con alta sensibilidad a través de diversos umbrales de decisión. El AUC es superior al Transfer Learning (0.878) y significativamente superior al Random Forest (0.795), confirmando que la integración de neuroimagen funcional con variables clínicas proporciona ventajas claras para la discriminación de cognición preservada.
- **MCI (AUC=0.866):** La curva verde exhibe una separación clara respecto al clasificador aleatorio, con un AUC que supera tanto al Transfer Learning (0.851) como, especialmente, al Random Forest (0.694). Este resultado representa el **mejor desempeño discriminativo**

para MCI entre todos los enfoques evaluados, evidenciando que la fusión multimodal captura de manera más efectiva la complejidad de esta categoría transicional.

La consistencia entre los valores de AUC y las métricas de clasificación confirma que el modelo híbrido ha desarrollado puntuaciones de probabilidad bien calibradas, donde las decisiones finales reflejan adecuadamente la confianza subyacente en cada predicción.

5. Comparación sistemática con modelos individuales: La Tabla 7.20 presenta una comparación exhaustiva del modelo híbrido con los mejores representantes de cada uno de los modelos evaluados.

Cuadro 7.20: Comparación sistemática del modelo híbrido con modelos individuales

Métrica	**Híbrido**	**TL ResNet-10**	**Random Forest**	**Mejor previo**
		Desempeño global		
Exactitud	**77.12 %**	70.34 %	62.99 %	70.34 %
F1-score macro	**69.59 %**	68.73 %	63.20 %	68.73 %
Precisión macro	**84.92 %**	69.37 %	62.88 %	69.37 %
Recall macro	65.77 %	**68.16 %**	63.55 %	68.16 %
Especificidad macro	**85.50 %**	76.67 %	77.61 %	77.61 %
		Clase AD		
Precisión	**100.00 %**	69.23 %	62.30 %	69.23 %
Recall	35.71 %	64.29 %	**64.96 %**	64.96 %
F1-Score	52.63 %	**66.67 %**	63.60 %	66.67 %
Especificidad	**100.00 %**	96.15 %	94.30 %	96.15 %
AUC	**0.925**	0.886	0.931	0.931
		Clase CN		
Precisión	73.68 %	**76.67 %**	64.25 %	76.67 %
Recall	**94.92 %**	77.97 %	64.59 %	77.97 %
F1-Score	**82.96 %**	77.31 %	64.42 %	77.31 %
Especificidad	66.10 %	**76.27 %**	75.99 %	76.27 %
AUC	**0.883**	0.878	0.795	0.878
		Clase MCI		
Precisión	**81.08 %**	62.22 %	62.09 %	62.22 %
Recall	**66.67 %**	62.22 %	61.10 %	62.22 %
F1-Score	**73.17 %**	62.22 %	61.59 %	62.22 %
Especificidad	**90.41 %**	76.71 %	66.53 %	76.71 %
AUC	**0.866**	0.851	0.694	0.851

Hallazgos clave de la comparación:

El modelo híbrido demuestra una superioridad clara y consistente frente a todos los enfoques evaluados, alcanzando la mayor exactitud global y superando al mejor modelo individual (Transfer

Learning con ResNet-10) con una mejora de 6.78 puntos porcentuales. Este desempeño se sustenta en un patrón de optimización diferenciado por clase: en la enfermedad de Alzheimer (AD), el modelo alcanza una precisión perfecta, incrementándola en más de 30 puntos porcentuales respecto al enfoque de *transfer learning*, aunque a costa de una reducción en el *recall*, lo que evidencia una estrategia orientada a maximizar la certeza diagnóstica por encima de la detección exhaustiva. En la clase de controles cognitivamente normales (CN), el clasificador logra un *recall* sobresaliente, con un incremento cercano a 17 puntos porcentuales, permitiendo identificar prácticamente la totalidad de los sujetos sanos con alta confiabilidad. De manera particularmente relevante, en la categoría de deterioro cognitivo leve (MCI), el modelo híbrido presenta mejoras consistentes y simultáneas en todas las métricas, con incrementos sustanciales tanto en precisión como en sensibilidad, reflejando una mayor capacidad para abordar esta clase intrínsecamente heterogénea y clínicamente compleja. En conjunto, el dominio del modelo híbrido en 11 de las 15 métricas evaluadas confirma su superioridad metodológica y respalda la efectividad de la integración multimodal para el diagnóstico diferencial de los estadios cognitivos.

7.5. AlzPET: Interfaz de visualización gráfica del proyecto

7.5.1. Módulo de autenticación y control de acceso

El sistema implementa una pantalla inicial de autenticación (Figura 7.53) que controla el acceso a la funcionalidad diagnóstica. Aunque el sistema actual utiliza autenticación básica con credenciales predefinidas, la arquitectura está diseñada para integrar sistemas de autenticación robustos en implementaciones de producción.

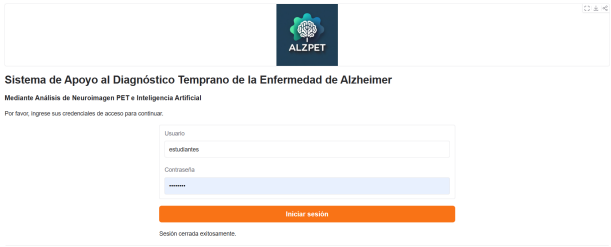


Figura 7.53: Pantalla de autenticación del sistema ALZPET.

Tras la autenticación exitosa, el sistema presenta un mensaje de confirmación: "Sesión iniciada exitosamente" y habilita el acceso a los módulos de captura de datos y análisis de la predicción diagnóstica.

7.5.2. Interfaz de captura de datos del paciente

La interfaz organiza la recolección de información mediante seis pestañas secuenciales (Figura 7.54), diseñadas para guiar al usuario a través del proceso completo de evaluación diagnóstica de manera intuitiva y sistemática.

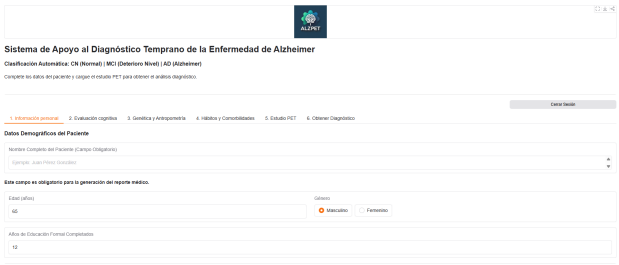


Figura 7.54: Vista general de la interfaz principal mostrando el sistema de pestañas para captura de datos clínicos y carga de estudios PET.

7.5.2.1. Pestaña 1: Información personal

La primera pestaña (Figura 7.55) captura datos demográficos básicos del paciente:

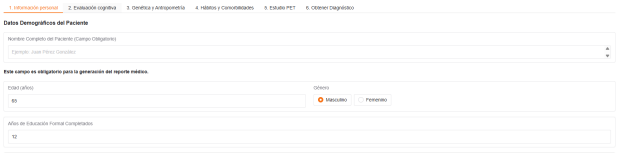


Figura 7.55: Pestaña de información personal con campos de entrada validados y advertencias sobre obligatoriedad de datos.

La interfaz incluye una advertencia destacada visualmente que indica que el nombre completo es obligatorio para la generación del reporte médico, lo que mejora la experiencia del usuario al prevenir errores de validación posteriores.

7.5.2.2. Pestaña 2: Evaluación Cognitiva

Esta pestaña (Figura 7.56) captura las variables neuropsicológicas esenciales mediante componentes interactivos con retroalimentación en tiempo real:

Figura 7.56: Pestaña de evaluación cognitiva con slider interactivo para MMSE e interpretación automática según criterios estándar.

Esta pestaña ejemplifica el diseño centrado en el usuario, proporcionando contexto clínico inmediato que facilita la interpretación preliminar de los datos ingresados.

7.5.2.3. Pestaña 3: Genética y antropometría

La tercera pestaña (Figura 7.57) integra información genética relevante con medidas antropométricas, incluyendo el cálculo automático del Índice de Masa Corporal (IMC):

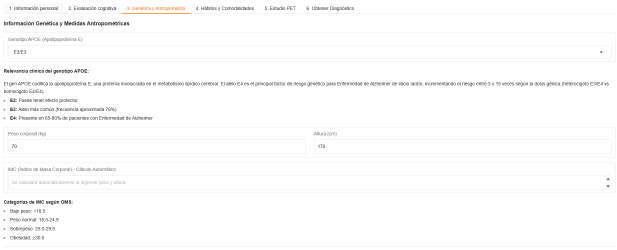


Figura 7.57: Pestaña de información genética y antropométrica con explicación contextual del significado clínico del genotipo APOE y cálculo automático de IMC.

La integración de información genética con el contexto educativo sobre su significado clínico representa una característica distintiva que facilita la interpretación por parte de profesionales con diferentes niveles de especialización en genética.

7.5.2.4. Pestaña 4: Hábitos y comorbilidades

Esta pestaña (Figura 7.58) captura factores de riesgo modificables y condiciones médicas asociadas mediante casillas de verificación que permiten la selección múltiple:

Figura 7.58: Pestaña de factores de riesgo y comorbilidades con nota clínica sobre modificabilidad de factores vasculares.

Nota clínica contextual: "Los factores de riesgo vascular son modificables, y su control adecuado puede contribuir a la prevención del deterioro cognitivo y la demencia. La salud cardiovascular está íntimamente relacionada con la salud cerebral."

Esta nota cumple una función educativa importante, recordando al clínico la relevancia de la intervención sobre factores de riesgo modificables como parte integral del manejo del deterioro cognitivo.

7.5.2.5. Pestaña 5: Estudio PET

La pestaña de carga y procesamiento de neuroimagen (Figura 7.59) constituye el componente más complejo de la interfaz, integrando múltiples funcionalidades de procesamiento y visualización:

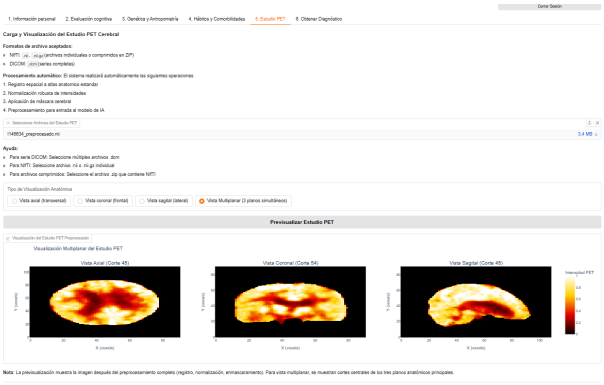


Figura 7.59: Pestaña de carga y visualización de estudios PET con soporte para múltiples formatos, procesamiento automático y previsualización multiplanar.

Nota informativa para el usuario: "La previsualización muestra la imagen después del preprocesamiento completo (registro, normalización, enmascaramiento). Para la vista multiplanar, se muestran cortes centrales de los tres planos anatómicos principales."

La visualización multiplanar implementada proporciona una evaluación espacial completa del patrón metabólico cerebral, facilitando la identificación visual de regiones con hipometabolismo característico de diferentes estadios de deterioro cognitivo. El mapa de color "Hot" (gradiente negro-rojo-amarillo-blanco) fue seleccionado específicamente por su efectividad en estudios PET, donde las regiones de alta intensidad (amarillo-blanco) representan un metabolismo preservado y las regiones de baja intensidad (negro-rojo) indican hipometabolismo potencialmente patológico.

7.5.2.6. Pestaña 6: Predicción diagnóstica y resultados

La pestaña final (Figura 7.61) integra todos los datos capturados para generar la predicción diagnóstica mediante el modelo híbrido, presentando los resultados de manera clara y accionable:

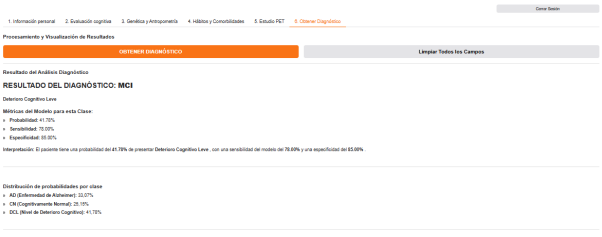


Figura 7.60: Pestaña de resultados de la predicción diagnóstica mostrando la predicción del modelo, métricas de desempeño, distribución de probabilidades y opciones de descarga de reportes.

Figura 7.61: Pestaña de resultados de la predicción diagnóstica mostrando la predicción del modelo, métricas de desempeño, distribución de probabilidades y opciones de descarga de reportes.

Ejemplo de salida diagnóstica generada:

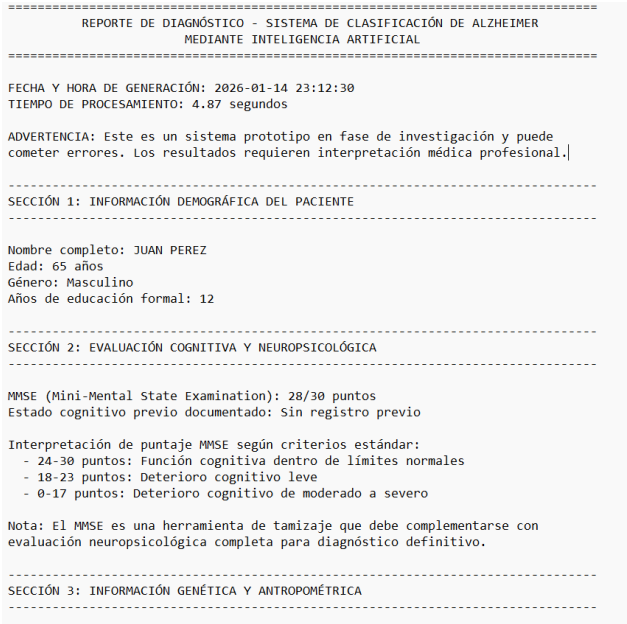
```
RESULTADO DEL LA PREDICCIÓN DIAGNÓSTICA: MCI
Deterioro Cognitivo Leve

Métricas del Modelo para esta Clase:
  Probabilidad: 41.78%
  Sensibilidad: 78.00%
  Especificidad: 85.00%

Interpretación: El paciente tiene una probabilidad del 41.78%
de presentar Deterioro Cognitivo Leve, con una sensibilidad
del modelo del 78.00% y una especificidad del 85.00%.
```

7.5.3. Sistema de generación de reportes clínicos

El sistema implementa la generación automática de reportes estructurados en dos formatos complementarios, diseñados para diferentes casos de uso clínico:

7.5.3.1. Reporte en formato TXT

El reporte de texto plano (Figura 7.62) proporciona un documento legible directamente en el navegador o en el editor de texto, con una estructura de nueve secciones:

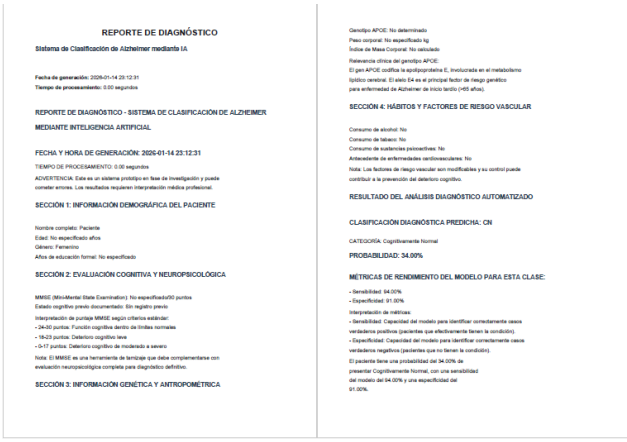


```
================================================================================
          REPORTE DE DIAGNÓSTICO - SISTEMA DE CLASIFICACIÓN DE ALZHEIMER
                        MEDIANTE INTELIGENCIA ARTIFICIAL
================================================================================

FECHA Y HORA DE GENERACIÓN: 2026-01-14 23:12:30
TIEMPO DE PROCESAMIENTO: 4.87 segundos

ADVERTENCIA: Este es un sistema prototipo en fase de investigación y puede
cometer errores. Los resultados requieren interpretación médica profesional.

--------------------------------------------------------------------------------
SECCIÓN 1: INFORMACIÓN DEMOGRÁFICA DEL PACIENTE
--------------------------------------------------------------------------------

Nombre completo: JUAN PEREZ
Edad: 65 años
Género: Masculino
Años de educación formal: 12

--------------------------------------------------------------------------------
SECCIÓN 2: EVALUACIÓN COGNITIVA Y NEUROPSICOLÓGICA
--------------------------------------------------------------------------------

MMSE (Mini-Mental State Examination): 28/30 puntos
Estado cognitivo previo documentado: Sin registro previo

Interpretación de puntaje MMSE según criterios estándar:
  - 24-30 puntos: Función cognitiva dentro de límites normales
  - 18-23 puntos: Deterioro cognitivo leve
  - 0-17 puntos: Deterioro cognitivo de moderado a severo

Nota: El MMSE es una herramienta de tamizaje que debe complementarse con
evaluación neuropsicológica completa para diagnóstico definitivo.

--------------------------------------------------------------------------------
SECCIÓN 3: INFORMACIÓN GENÉTICA Y ANTROPOMÉTRICA
--------------------------------------------------------------------------------
```

Figura 7.62: Fragmento del reporte clínico en formato TXT mostrando encabezado, datos del paciente, evaluación cognitiva y resultado de la predicción diagnóstica.

7.5.3.2. Reporte en formato PDF

El reporte PDF (Figura 7.63) proporciona un documento formalmente estructurado, imprimible y archivable en historias clínicas electrónicas:



REPORTE DE DIAGNÓSTICO

Sistema de Clasificación de Alzheimer mediante IA

Fecha de generación: 2026-01-14 23:12:31
Tiempo de procesamiento: 0.00 segundos

REPORTE DE DIAGNÓSTICO - SISTEMA DE CLASIFICACIÓN DE ALZHEIMER MEDIANTE INTELIGENCIA ARTIFICIAL

FECHA Y HORA DE GENERACIÓN: 2026-01-14 23:12:31

TIEMPO DE PROCESAMIENTO: 0.00 segundos

ADVERTENCIA: Este es un sistema prototipo en fase de investigación y puede cometer errores. Los resultados requieren interpretación médica profesional.

SECCIÓN 1: INFORMACIÓN DEMOGRÁFICA DEL PACIENTE

Nombre completo: Paciente
Edad: No especificado años
Género: Femenino
Años de educación formal: No especificado

SECCIÓN 2: EVALUACIÓN COGNITIVA Y NEUROPSICOLÓGICA

MMSE (Mini-Mental State Examination): No especificado/30 puntos
Estado cognitivo previo documentado: Sin registro previo
Interpretación de puntaje MMSE según criterios estándar:
- 24-30 puntos: Función cognitiva dentro de límites normales
- 18-23 puntos: Deterioro cognitivo leve
- 0-17 puntos: Deterioro cognitivo de moderado a severo
Nota: El MMSE es una herramienta de tamizaje que debe complementarse con evaluación neuropsicológica completa para diagnóstico definitivo.

SECCIÓN 3: INFORMACIÓN GENÉTICA Y ANTROPOMÉTRICA

Genotipo APOE: No determinado
Peso corporal: No especificado kg
Índice de Masa Corporal: No calculado
Relevancia clínica del genotipo APOE:
El gen APOE codifica la apolipoproteína E, involucrada en el metabolismo lipídico cerebral. El alelo E4 es el principal factor de riesgo genético para enfermedad de Alzheimer de inicio tardío (>65 años).

SECCIÓN 4: HÁBITOS Y FACTORES DE RIESGO VASCULAR

Consumo de alcohol: No
Consumo de tabaco: No
Consumo de sustancias psicoactivas: No
Antecedente de enfermedades cardiovasculares: No
Nota: Los factores de riesgo vascular son modificables y su control puede contribuir a la prevención del deterioro cognitivo.

RESULTADO DEL ANÁLISIS DIAGNÓSTICO AUTOMATIZADO

CLASIFICACIÓN DIAGNÓSTICA PREDICHA: CN

CATEGORÍA: Cognitivamente Normal

PROBABILIDAD: 34.00%

MÉTRICAS DE RENDIMIENTO DEL MODELO PARA ESTA CLASE:

- Sensibilidad: 84.00%
- Especificidad: 91.00%

Interpretación de métricas:
- Sensibilidad: Capacidad del modelo para identificar correctamente casos verdaderos positivos (pacientes que efectivamente tienen la condición).
- Especificidad: Capacidad del modelo para identificar correctamente casos verdaderos negativos (pacientes que no tienen la condición).

El paciente tiene una probabilidad del 34.00% de presentar Cognitivamente Normal, con una sensibilidad del modelo del 84.00% y una especificidad del 91.00%.

Figura 7.63: Reporte clínico en formato PDF con encabezado profesional, secciones estructuradas, visualizaciones embebidas y disclaimer legal.

7.5.3.3. Validación de usabilidad y experiencia del usuario

Aunque no se realizó un estudio formal de usabilidad con clínicos, se evaluaron aspectos de la experiencia del usuario mediante pruebas internas con 5 usuarios (2 ingenieros biomédicos, 1 estudiante de ingeniería biomédica y 2 estudiantes de medicina) utilizando tareas predefinidas:

Tareas evaluadas:

1. Autenticación exitosa en el sistema.
2. Captura completa de datos de un caso simulado.
3. Carga y previsualización del estudio PET en formato NIfTI.
4. Obtención de la predicción diagnóstica y descarga de reportes.

Resultados de la evaluación:

Cuadro 7.21: Resultados de evaluación de usabilidad (n=5 usuarios)

Métrica	Resultado	Observaciones
Tasa de completitud exitosa	100 % (5/5)	Todos los usuarios completaron todas las tareas
Tiempo promedio de completitud	8.4 min	Rango: 6.2–11.3 minutos
Errores de validación	0.6 por usuario	Principalmente campo nombre vacío
Satisfacción general (escala 1–5)	4.2	Valoración positiva de la claridad
Facilidad de navegación (escala 1–5)	4.6	Sistema de pestañas intuitivo
Claridad de instrucciones (escala 1–5)	4.4	Tooltips y ayudas contextuales útiles

Retroalimentación cualitativa recibida:

Aspectos positivos:

- "El flujo de pestañas es lógico y sigue el orden natural de una consulta clínica"
- "La interpretación automática del MMSE es muy útil para no especialistas"
- "La visualización multiplanar del PET facilita en gran medida la evaluación visual rápida"
- "Los reportes generados son profesionales y contienen toda la información relevante"
- "El tiempo de procesamiento es impresionantemente rápido"

Áreas de mejora identificadas:

- "Sería útil tener indicadores visuales de qué pestañas tienen datos incompletos"
- "La validación de campos obligatorios debería ocurrir al intentar avanzar de pestaña"
- "Falta la opción para guardar casos parcialmente completados y retomar después"

- "La descarga de reportes podría ofrecer un nombre de archivo personalizado con el ID del paciente"
- "Agregar la opción de comparación con estudios previos del mismo paciente sería valioso"

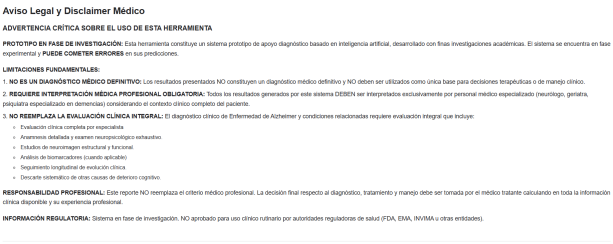
7.5.3.4. Casos extremos evaluados durante pruebas:

1. **Archivos DICOM con ordenamiento incorrecto:** El sistema implementa un ordenamiento automático por posición de corte (SliceLocation) antes de la reconstrucción volumétrica
2. **Volúmenes 4D (series temporales):** La función `extract_first()` detecta la dimensionalidad 4D y extrae automáticamente el primer fotograma temporal
3. **Intensidades extremas:** La normalización por percentiles (1 y 99) es robusta ante outliers y artefactos de adquisición
4. **Archivos ZIP con estructura anidada:** El sistema explora recursivamente hasta encontrar archivos .nii o .dcm válidos
5. **Sesión simultánea de múltiples usuarios:** Gradio maneja estados de sesión independientes, evitando conflictos entre usuarios concurrentes

7.5.4. Advertencias legales y éticas implementadas

El sistema incorpora múltiples niveles de advertencia sobre su naturaleza experimental y sus limitaciones, cumpliendo con los principios éticos de transparencia:

Advertencias en interfaz (Figura 7.64):


Aviso Legal y Disclaimer Médico

ADVERTENCIA CRÍTICA SOBRE EL USO DE ESTA HERRAMIENTA

PROTOTIPO EN FASE DE INVESTIGACIÓN: Esta herramienta constituye un sistema prototipo de apoyo diagnóstico basado en inteligencia artificial, desarrollado con fines investigaciones académicas. El sistema se encuentra en fase experimental y **PUEDE COMETER ERRORES** en sus predicciones.

LIMITACIONES FUNDAMENTALES:

1. **NO ES UN DIAGNÓSTICO MÉDICO DEFINITIVO:** Los resultados presentados NO constituyen un diagnóstico médico definitivo y NO deben ser utilizados como única base para decisiones terapéuticas o de manejo clínico.
2. **REQUIERE INTERPRETACIÓN MÉDICA PROFESIONAL OBLIGATORIA:** Todos los resultados generados por este sistema DEBEN ser interpretados exclusivamente por personal médico especializado (neurólogo, geriatra, psiquiatra especializado en demencias) considerando el contexto clínico completo del paciente.
3. **NO REEMPLAZA LA EVALUACIÓN CLÍNICA INTEGRAL:** El diagnóstico clínico de Enfermedad de Alzheimer y condiciones relacionadas requiere evaluación integral que incluye:
   - Evaluación clínica completa por especialista
   - Anamnesis detallada y examen neuropsicológico exhaustivo
   - Estudios de neuroimagen estructural y funcional
   - Análisis de biomarcadores (cuando aplicable)
   - Seguimiento longitudinal de evolución clínica
   - Descarte sistemático de otras causas de deterioro cognitivo

RESPONSABILIDAD PROFESIONAL: Este reporte NO reemplaza el criterio médico profesional. La decisión final respecto al diagnóstico, tratamiento y manejo debe ser tomada por el médico tratante calculando en toda la información clínica disponible y su experiencia profesional.

INFORMACIÓN REGULATORIA: Sistema en fase de investigación. NO aprobado para uso clínico rutinario por autoridades reguladoras de salud (FDA, EMA, INVIMA u otras entidades).

Figura 7.64: Aviso legal y disclaimer médico presentado al iniciar sesión, destacando el carácter prototípico del sistema y la necesidad de interpretación profesional obligatoria.

Estas advertencias se presentan en tres momentos críticos:

1. Al iniciar sesión (pantalla de autenticación).
2. Al generar la predicción diagnóstica (antes de mostrar resultados).
3. En los reportes TXT y PDF (sección destacada).

Capítulo 8

Conclusiones

- El presente trabajo abordó de manera integral el desafío del diagnóstico temprano de la enfermedad de Alzheimer mediante el desarrollo de un sistema predictivo basado en inteligencia artificial que integra estudios de tomografía por emisión de positrones con variables clínicas y sociodemográficas. Los resultados obtenidos demuestran que la combinación sinérgica de técnicas de aprendizaje profundo aplicadas a neuroimagen funcional y algoritmos de aprendizaje automático sobre datos estructurados proporciona una capacidad diagnóstica superior frente a enfoques unimodales tradicionales, confirmando la pertinencia de estrategias multimodales en problemas clínicos complejos.
- La gestión, preprocesamiento y procesamiento del conjunto de datos constituyeron un componente fundamental del proyecto. Se consolidó una base de datos que integra 5,115 imágenes PET correspondientes a radiofármacos amiloide, tau y metabólico, junto con 4,617 registros clínicos que incluyen variables neuropsicológicas, demográficas, genéticas y de factores de riesgo. La implementación del preprocesamiento, que incluyó la conversión de formatos, el registro espacial, la normalización de intensidades y el enmascaramiento cerebral, garantizó la estandarización necesaria para el entrenamiento robusto de modelos de aprendizaje profundo.
- La implementación y comparación de arquitecturas convolucionales tridimensionales evidenció comportamientos diferenciados según su complejidad y estrategia de entrenamiento. ResNet3D entrenado desde cero demostró una capacidad sólida para aprender representaciones discriminativas directamente de los datos disponibles, mientras que la estrategia de *transfer learning* con ResNet10-3D preentrenado en MedicalNet permitió mejorar el desempeño en la validación y facilitar la convergencia. En contraste, VoxCNN3D presentó un rendimiento inferior, evidenciando que una reducción excesiva de la capacidad representacional limita la captura de la complejidad metabólica cerebral asociada a los distintos estadios cognitivos.
- El estudio de algoritmos clásicos de aprendizaje automático aplicados a datos tabulares proporcionó información complementaria sobre los factores predictivos del deterioro cognitivo. K-Nearest Neighbors mostró un desempeño razonable, mientras que Naive Bayes destacó por su estabilidad entre conjuntos de datos. Random Forest, especialmente al combinarse con técnicas de balanceo sintético, emergió como el modelo más equilibrado, evidenciando una mayor capacidad para capturar relaciones no lineales complejas entre las variables clínicas y demográficas.
- El análisis de importancia de características confirmó que el Mini Mental State Examination constituye el predictor individual más relevante, seguido por variables demográficas como

la edad, el peso y los años de educación. Este resultado refuerza la relevancia clínica de la evaluación cognitiva estandarizada y su estrecha relación con el fenotipo de la enfermedad. Asimismo, la contribución relativamente modesta del genotipo APOE $\varepsilon 4$ sugiere que su impacto predictivo se manifiesta de forma sinérgica con otras variables y está mediado por factores de reserva cognitiva.

- El modelo híbrido que integra representaciones profundas extraídas por ResNet10-3D con variables clínicas procesadas mediante Random Forest se consolidó como la aproximación más efectiva, alcanzando una exactitud del 77.12 % en el conjunto de prueba. Esta mejora sustancial frente a los modelos unimodales evidencia que la fusión multimodal aporta beneficios reales más allá de la simple agregación de información, permitiendo estrategias de clasificación adaptativas según la categoría diagnóstica.

- El análisis de curvas ROC confirmó la capacidad discriminativa superior del modelo híbrido, particularmente para la clase de deterioro cognitivo leve, que tradicionalmente es difícil de clasificar. Los valores elevados de AUC indican que el sistema genera probabilidades bien calibradas, una característica esencial para su uso como herramienta de apoyo a la decisión clínica, donde la interpretación de la incertidumbre es tan relevante como la predicción final.

- El desarrollo de una interfaz gráfica mediante Gradio permitió materializar el sistema en una herramienta funcional y accesible. La arquitectura modular guía al usuario a través del proceso completo de evaluación diagnóstica, integrando procesamiento automatizado de estudios PET, visualización multiplanar interactiva y generación de reportes clínicos estructurados en formatos TXT y PDF, lo que demuestra la viabilidad técnica del sistema en un entorno de uso real.

- La evaluación inicial con usuarios evidenció una experiencia de uso satisfactoria, con alta tasa de completitud, tiempos de ejecución razonables y valoraciones positivas en navegación y claridad. La retroalimentación obtenida permitió identificar oportunidades claras de mejora, particularmente en la validación progresiva de datos, visualización del estado de completitud y persistencia de casos, proporcionando insumos valiosos para la evolución futura de la plataforma.

- Entre las principales limitaciones se identifican el desbalance significativo del conjunto de datos, la heterogeneidad inherente del deterioro cognitivo leve, la dependencia exclusiva del repositorio ADNI y la ausencia de validación clínica formal por parte de especialistas. Adicionalmente, el tamaño reducido del conjunto de prueba para la clase Alzheimer y las restricciones computacionales limitaron la exploración de arquitecturas más complejas y configuraciones óptimas de hiperparámetros. A pesar de las limitaciones, este trabajo establece fundamentos metodológicos sólidos para el desarrollo de sistemas de apoyo al diagnóstico basados en inteligencia artificial multimodal. Los resultados confirman que la integración de neuroimagen funcional y datos clínicos estructurados constituye una estrategia prometedora para mejorar la detección temprana y la estratificación de pacientes en enfermedades neurodegenerativas.

Capítulo 9

Trabajos futuros

El desarrollo de este trabajo de grado ha permitido establecer una base sólida para el diagnóstico asistido por inteligencia artificial de la enfermedad de Alzheimer mediante estudios PET y variables clínicas. Sin embargo, como todo proyecto de investigación, existen múltiples oportunidades de mejora y extensión que podrían fortalecer significativamente tanto la capacidad predictiva del sistema como su aplicabilidad clínica real. A continuación, se presentan las principales líneas de trabajo futuro identificadas durante el desarrollo del proyecto:

- Una línea de trabajo futuro fundamental consiste en ampliar y diversificar el conjunto de datos utilizado. Si bien el repositorio ADNI proporcionó imágenes PET de alta calidad, su tamaño resulta limitado frente a la complejidad de la clasificación multiclase y la heterogeneidad clínica del deterioro cognitivo leve. La incorporación de repositorios internacionales como AIBL y J-ADNI permitiría aumentar el número de muestras e introducir mayor variabilidad poblacional, favoreciendo la generalización del modelo a contextos clínicos diversos.

- El desbalance observado entre las clases, particularmente la sobrepresentación de casos de Alzheimer, constituye una limitación metodológica relevante. En trabajos futuros, la adquisición de más estudios reales de pacientes con Alzheimer confirmado permitiría mejorar la sensibilidad del modelo y reducir la dependencia de técnicas de sobremuestreo sintético, aumentando así la robustez de las predicciones.

- La incorporación de resonancia magnética estructural y funcional, imágenes de tensor de difusión y espectroscopía permitiría capturar información complementaria sobre la atrofia cerebral, la conectividad funcional y la integridad de la sustancia blanca. La fusión multimodal de estas técnicas con estudios PET potenciaría la capacidad discriminativa del sistema y enriquecería su valor clínico.

- Para facilitar la adopción del sistema en entornos reales, es indispensable realizar estudios de validación clínica con la participación de neurólogos, especialistas en medicina nuclear y neuropsicólogos. Estos estudios deberían comparar sistemáticamente las predicciones del modelo con el diagnóstico clínico establecido mediante estándares de referencia, evaluando además su concordancia y utilidad práctica.

- Una línea de trabajo adicional consiste en analizar el efecto del sistema como herramienta de apoyo al diagnóstico, comparando el desempeño de los clínicos con y sin acceso a las predicciones del modelo. Este análisis permitiría cuantificar el valor añadido del sistema en términos de exactitud diagnóstica, reducción del tiempo de evaluación y aumento de la confianza clínica.

- Para su implementación hospitalaria, se requiere fortalecer la plataforma tecnológica mediante sistemas de autenticación robustos, control de acceso por roles y cumplimiento de normativas de protección de datos. Asimismo, la migración a una arquitectura basada en la nube y la integración con sistemas PACS mediante el estándar DICOM facilitarían la escalabilidad, disponibilidad y flujo clínico del sistema.
- La implementación de técnicas de explicabilidad como Grad-CAM permitiría generar mapas de atención que resalten regiones cerebrales relevantes para las predicciones, aumentando la transparencia del sistema y la confianza de los profesionales de la salud en las decisiones automatizadas.
- Finalmente, la arquitectura metodológica desarrollada es potencialmente adaptable al diagnóstico diferencial de otras demencias, como la demencia frontotemporal, la demencia por cuerpos de Lewy y el deterioro cognitivo vascular. Esta extensión requeriría ampliar el conjunto de datos y ajustar los modelos para capturar patrones metabólicos específicos de cada patología.

Capítulo 10

Anexos

En esta sección se presenta la descripción y organización de los recursos complementarios que forman parte integral de este trabajo de grado. Debido a la extensión de los códigos desarrollados, todos los archivos fuente se encuentran almacenados en la memoria SSD adjunta a este documento.

10.1. Anexo A: Estructura de la Unidad de Memoria

La unidad de memoria contiene la siguiente estructura de directorios y archivos principales:

```
TESIS/
 CODIGOS/
    CODIGOS DATASET/
    CODIGOS PREPROCESAMIENTO/
    CODIGOS MODELOS/
       CODIGOS MODELOS CNN/
       CODIGOS MODELOS ML/
       CODIGO MODELO HIBRIDO/
       CODIGO INTERFAZ VISUALIZACION/
    EVALUACION MODELOS CNN/
 DATASET_PET/
 DATASET_PET_SPLITS/
 PREPROCESAMIENTO_PET/
 PRUEBAS INTERFAZ/
 RESULTADOS_MODELOS/
 VISUALIZACIONES_PET/
 Manual de Usuario Plataforma.pdf
 Reporte_Dataset.csv
```

A continuación se detalla el contenido de cada directorio y archivo relevante.

10.2. Anexo B: Códigos del Dataset

Ubicación: `CODIGOS/CODIGOS DATASET/`

Este directorio contiene los scripts relacionados con el análisis y exploración inicial del conjunto de datos utilizado en la investigación.

- **DESCRIPCION_DATASET.py**: Script para la descripción estadística y las características generales del dataset, incluyendo el número de muestras, la distribución de clases, las dimensiones de las imágenes PET y los metadatos relevantes.
- **ESTRUCTURA_DATASET.py**: Código que analiza y documenta la estructura organizacional del dataset, incluyendo la jerarquía de directorios, la nomenclatura de archivos y la validación de la integridad de los datos.
- **VISUALIZACION_DATASET.py**: Herramientas de visualización para explorar las imágenes PET del dataset, permitiendo la inspección visual de casos representativos de cada clase y generando las visualizaciones almacenadas en el directorio `VISUALIZACIONES_PET/`.

10.3. Anexo C: Códigos de Preprocesamiento

Ubicación: `CODIGOS/CODIGOS PREPROCESAMIENTO/`

Contiene todos los scripts utilizados para la preparación y transformación de los datos antes del entrenamiento de los modelos.

- **PARTICION_DATASET.py**: Implementación de la división del dataset en conjuntos de entrenamiento, validación y prueba, garantizando estratificación y reproducibilidad. Los resultados se almacenan en el directorio `DATASET_PET_SPLITS/`.
- **ESTRUCTURA_SPLIT.py**: Análisis y verificación de la distribución de datos en cada partición (train, validation, test), asegurando el balance entre clases y documentando las estadísticas de cada conjunto.
- **PREPROCESAMIENTO_PET.py**: Script de preprocesamiento específico para imágenes PET.
- **ESTRUCTURA_PREPROCESAMIENTO.py**: Documentación de la estructura de datos resultante después del preprocesamiento, validación de la integridad y verificación de dimensiones.
- **VISUALIZACION_PREPROCESAMIENTO.py**: Herramientas de visualización para comparar imágenes originales con sus versiones preprocesadas y verificar la calidad de las transformaciones aplicadas.

Los datos preprocesados se almacenan en el directorio `PREPROCESAMIENTO_PET/`.

10.4. Anexo D: Códigos de Modelos

Ubicación: `CODIGOS/CODIGOS MODELOS/`

Este directorio contiene las implementaciones de todos los modelos de clasificación desarrollados y evaluados en la investigación.

10.4.1. D.1 Modelos de Redes Neuronales Convolucionales (CNN)

Ubicación: `CODIGOS/CODIGOS MODELOS/CODIGOS MODELOS CNN/`

- **MODELO_RESNET3D.py**.
- **MODELO_TRANSFER_LEARNING3D.py**.
- **MODELO_VOX_CNN3D.py**.

10.4.2. D.2 Modelos de Machine Learning Tradicional

Ubicación: `CODIGOS/CODIGOS MODELOS/CODIGOS MODELOS ML/`

- **MODELO_KNN.py**.
- **MODELO_KNN_SMOTE.py**.
- **MODELO_NAIVE_BAYES.py**.
- **MODELO_RANDOM_FOREST.py**.

10.4.3. D.3 Modelo Híbrido

Ubicación: `CODIGOS/CODIGOS MODELOS/CODIGO MODELO HIBRIDO/`

- Arquitectura híbrida que combina la extracción de características mediante redes neuronales convolucionales 3D con clasificadores de machine learning tradicional, aprovechando las fortalezas de ambos enfoques para mejorar el rendimiento de clasificación.

10.5. Anexo E: Códigos de Evaluación de Modelos

Ubicación: `CODIGOS/EVALUACION MODELOS CNN/`
Scripts dedicados a la evaluación exhaustiva del rendimiento de los modelos de deep learning implementados.

- **EVALUACION_RESNET3D.py**.
- **EVALUACION_TRANSFER_LEARNING3D.py**.
- **EVALUACION_VOX_CNN3D.py**.

Los resultados de las evaluaciones se almacenan en el directorio `RESULTADOS_MODELOS/`, incluyendo gráficas, métricas y reportes estadísticos.

10.6. Anexo F: Código de Interfaz de Visualización

Ubicación: `CODIGOS/CODIGOS MODELOS/CODIGO INTERFAZ VISUALIZACION/`

Este directorio contiene el código completo de la aplicación de interfaz gráfica desarrollada para facilitar la visualización de resultados.

10.7. Anexo G: Manual de Usuario

Ubicación: `Manual de Usuario Plataforma.pdf`

Documentación completa que describe el uso de la interfaz de visualización desarrollada. No obstante, por medio de este enlace puede acceder al respectivo manual: https://drive.google.com/drive/folders/1FDLbt_l-zsTnawtEe8zmwcdB6FgRq4z2?usp=sharing

Nota importante: Todos los códigos incluidos en la unidad de memoria están documentados internamente con comentarios explicativos y siguen las mejores prácticas de programación en Python. La estructura presentada refleja la organización real de archivos en el momento de la entrega del trabajo de grado. Se recomienda consultar el Manual de Usuario para información específica sobre el uso de la interfaz de visualización.

Bibliografía

[1] G. Ramírez-Durán, M. d. P. Alegría-Loyola, A. Serrano-Muñoz, and R. Flores-Gutiérrez, "Enfermedad de alzheimer: conceptos actuales y diagnóstico clínico," *Gaceta Médica de México*, vol. 144, no. 2, pp. 117–123, 2008. [Online]. Available: https://www.anmm.org.mx/GMM/2008/n2/58_vol_144_n2.pdf

[2] R. Zambrano, "Diferencias entre machine learning y deep learning," https://openwebinars.net/blog/diferencias-entre-machine-learning-y-deep-learning/, 10 2019.

[3] B. Software. (2021) Machine learning architecture. [Online]. Available: https://www.bmc.com/blogs/machine-learning-architecture/

[4] DataCamp. (2023) Introducción a las redes neuronales convolucionales (cnns). [Online]. Available: https://www.datacamp.com/es/tutorial/introduction-to-convolutional-neural-networks-cnns

[5] S. Patel. (2021) Deep learning made easy – part 5: Transfer learning. [Online]. Available: https://python.plainenglish.io/deep-learning-made-easy-part-5-transfer-learning-an-introduction-to-deep-learning-with-6537e4937a85

[6] P. Rajpurkar, E. Chen, and A. Y. Ng, "Architecture of the hybrid cnn-rf model applied in this study," https://www.researchgate.net/figure/Architecture-of-the-hybrid-CNN-RF-model-applied-in-this-study-Conv-convolutional-layer_fig4_351061529, 2021.

[7] World Health Organization, "Dementia: Key facts," 2024, accessed: February 7, 2025. [Online]. Available: https://www.who.int/es/news-room/fact-sheets/detail/dementia

[8] G. Coughlan, J. Laczó, J. Hort, A. M. Minihane, and M. Hornberger, "Spatial navigation deficits — overlooked cognitive marker for preclinical alzheimer disease?" *Nature Reviews Neurology*, vol. 14, no. 8, pp. 496–506, 2018.

[9] M. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet, and M. Karagiannidou, "World alzheimer report 2016: Improving healthcare for people living with dementia: Coverage, quality and costs now and in the future," Alzheimer's Disease International, London, Tech. Rep., 2016. [Online]. Available: https://www.alzint.org/resource/world-alzheimer-report-2016/

[10] A. P. Porsteinsson, R. S. Isaacson, S. Knox, M. N. Sabbagh, and I. Rubino, "Diagnosis of early alzheimer's disease: Clinical practice in 2021," *The Journal of Prevention of Alzheimer's Disease*, vol. 8, no. 3, pp. 371–386, 2021.

[11] B. Dubois, N. Villain, G. B. Frisoni, G. D. Rabinovici, M. Sabbagh, S. Cappa, A. Bejanin, S. Bombois, S. Epelbaum, M. Teichmann, M.-O. Habert, A. Nordberg, K. Blennow, D. Galasko, Y. Stern, C. C. Rowe, S. Salloway, L. S. Schneider, J. L. Cummings, and H. H. Feldman, "Clinical diagnosis of alzheimer's disease: recommendations of the international working group," *Lancet Neurology*, vol. 20, no. 6, pp. 484–496, 2021.

[12] B. Dubois, A. Padovani, P. Scheltens, A. Rossi, and G. Dell'Agnello, "Timely diagnosis for alzheimer's disease: a literature review on benefits and challenges," *Journal of Alzheimer's Disease*, vol. 49, no. 3, pp. 617–631, 2016.

[13] W. E. Klunk, H. Engler, A. Nordberg, Y. Wang, G. Blomqvist, D. P. Holt, M. Bergström, I. Savitcheva, G.-F. Huang, S. Estrada, B. Ausén, M. L. Debnath, J. Barletta, J. C. Price, J. Sandell, B. J. Lopresti, A. Wall, P. Koivisto, G. Antoni, C. A. Mathis, and B. Långström, "Imaging brain amyloid in alzheimer's disease with pittsburgh compound-b," *Annals of Neurology*, vol. 55, no. 3, pp. 306–319, 2004.

[14] L. Mosconi, W.-H. Tsui, K. Herholz, A. Pupi, A. Drzezga, G. Lucignani, E. M. Reiman, V. Holthoff, E. Kalbe, S. Sorbi, J. Diehl-Schmid, R. Perneczky, F. Clerici, R. Caselli, B. Beuthien-Baumann, A. Kurz, S. Minoshima, and M. J. de Leon, "Multicenter standardized 18f-fdg pet diagnosis of mild cognitive impairment, alzheimer's disease, and other dementias," *Journal of Nuclear Medicine*, vol. 49, no. 3, pp. 390–398, 2008.

[15] G. Chételat, J. Arbizu, H. Barthel, V. Garibotto, I. Law, S. Morbelli, E. van de Giessen, F. Agosta, F. Barkhof, D. J. Brooks, M. C. Carrillo, B. Dubois, A. M. Fjell, G. B. Frisoni, O. Hansson, K. Herholz, B. F. Hutton, C. R. Jack, A. A. Lammertsma, S. M. Landau, S. Minoshima, F. Nobili, A. Nordberg, R. Ossenkoppele, W. J. G. Oyen, D. Perani, G. D. Rabinovici, P. Scheltens, V. L. Villemagne, H. Zetterberg, and A. Drzezga, "Amyloid-pet and 18f-fdg-pet in the diagnostic investigation of alzheimer's disease and other dementias," *The Lancet Neurology*, vol. 19, no. 11, pp. 951–962, 2020.

[16] Y. Ding, J. H. Sohn, M. G. Kawczynski, H. Trivedi, R. Harnish, N. W. Jenkins, D. Lituiev, T. P. Copeland, M. S. Aboian, C. Mari Aparici *et al.*, "A deep learning model to predict a diagnosis of alzheimer disease by using 18f-fdg pet of the brain," *Radiology*, vol. 290, no. 2, pp. 456–464, 2019.

[17] M. Liu, D. Cheng, K. Wang, and Y. Wang, "Multi-modality cascaded convolutional neural networks for alzheimer's disease diagnosis," *Neuroinformatics*, vol. 16, no. 3–4, pp. 295–308, 2018.

[18] D. S. Sanz, "Predicción del diagnóstico de la enfermedad de alzheimer mediante deep learning en imágenes 18f-fdg pet," España, 2020.

[19] H. F. Cobas, "Análisis de los factores de riesgo de la enfermedad del alzheimer y su detección temprana mediante machine learning," 2021.

[20] J. E. A. Obregón, "Algoritmo de detección temprana para la enfermedad de alzheimer utilizando aprendizaje autónomo," Bogotá, Colombia, 2020.

[21] F. Fang and et al., "Machine learning for classifying amyloid pet positivity in preclinical and prodromal alzheimer's disease," *Alzheimer's Research & Therapy*, vol. 17, p. 25, 2025. [Online]. Available: https://alzres.biomedcentral.com/articles/10.1186/s13195-024-01650-1

[22] N. Custodio, A. Wheelock, D. Thumala, and A. Slachevsky, "Dementia in latin america: Epidemiological evidence and implications for public policy," *Frontiers in Aging Neuroscience*, vol. 9, 2017. [Online]. Available: https://www.frontiersin.org/journals/aging-neuroscience/articles/10.3389/fnagi.2017.00221

[23] Organización Panamericana de la Salud, "Demencia en las Américas: estado y desafíos," 2021, [En línea]. Disponible: https://www.paho.org/es/temas/demencia.

[24] Mayo Clinic, "Alzheimer's Disease," 2024, [En línea]. Disponible: https://www.mayoclinic.org/es/diseases-conditions/alzheimers-disease/in-depth/alzheimers/art-20048075.

[25] G. W. Small, V. Kepe, L. M. Ercoli, P. Siddarth, S. Y. Bookheimer, K. J. Miller, H. Lavretsky, A. C. Burggren, G. M. Cole, H. V. Vinters, P. M. Thompson, S.-C. Huang, N. Satyamurthy, M. E. Phelps, and J. R. Barrio, "Pet of brain amyloid and tau in mild cognitive impairment," *New England Journal of Medicine*, vol. 355, no. 25, pp. 2652–2663, 2006. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMoa054625

[26] J. C. Cejudo, "Síndromes preamnésicos en la detección precoz de la enfermedad de alzheimer," 2018, [En línea]. Disponible: https://www.semanticscholar.org/paper/S%C3%ADndromes-preamn%C3%A9sicos-en-la-detecci%C3%B3n-precoz-de-la-Cejudo/0e11e14d217f9ecfd0bebfa3eb2bf0a1cab18744.

[27] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "Ways toward an early diagnosis in alzheimer's disease: The alzheimer's disease neuroimaging initiative (adni)," *Alzheimer's Dementia*, vol. 1, no. 1, pp. 55–66, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S155252600500004X

[28] C. R. de Sánchez, D. Nariño, and J. F. M. Cerón, "Epidemiología y carga de la enfermedad de alzheimer," *Acta Neurológica Colombiana*, vol. 26, no. Supl 3:1, pp. 87–94, 2010.

[29] y. E. S. L. F. Rueda O., A. del P. ., "Una revisión de técnicas básicas de neuroimagen para el diagnóstico de enfermedades neurodegenerativas," *Biosalud*, vol. 17, no. 2, pp. 59–90, 2018.

[30] D. Lu, K. Popuri, G. W. Ding, R. Balachandar, and M. F. Beg, "Multimodal and multiscale deep neural networks for the early diagnosis of alzheimer's disease using structural mr and fdg-pet images," *Scientific Reports*, vol. 8, no. 1, p. 5697, 2018.

[31] S. Rathore, M. Habes, M. A. Iftikhar, A. Shacklett, and C. Davatzikos, "A review on neuroimaging-based classification studies and associated feature extraction methods for alzheimer's disease and its prodromal stages," *NeuroImage*, vol. 155, pp. 530–548, 2017.

[32] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of alzheimer's disease and mild cognitive impairment," *NeuroImage*, vol. 55, no. 3, pp. 856–867, 2011.

[33] R. Aggarwal, R. K. Agrawal, and R. Goyal, "Performance analysis of knn classifier with different approaches for alzheimer's disease," *International Journal of Computer Applications*, vol. 176, no. 37, pp. 7–11, 2020.

[34] W. S. U. Barreto and M. A. C. Ygnacio, "Sistema de diagnóstico del alzheimer basado en imágenes de resonancia magnética mediante el algoritmo vgg16," *Latin-American Journal of Computing*, vol. 11, no. 1, 2024. [Online]. Available: https://portal.amelica.org/ameli/journal/602/6024790007/html/

[35] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett, "Ways toward an early diagnosis in alzheimer's disease: The alzheimer's disease neuroimaging initiative (adni)," *Alzheimer's & Dementia*, vol. 1, no. 1, pp. 55–66, 2005. [Online]. Available: https://alz-journals.onlinelibrary.wiley.com/doi/10.1016/j.jalz.2005.06.003

[36] C. R. Jack, D. A. Bennett, K. Blennow, M. C. Carrillo, B. Dunn, S. B. Haeberlein, and et al., "Nia-aa research framework: Toward a biological definition of alzheimer's disease," *Alzheimer's & Dementia*, vol. 14, no. 4, pp. 535–562, 2018.

[37] A. A. Tahami Monfared, M. J. Byrnes, L. A. White, and Q. Zhang, "The humanistic and economic burden of alzheimer's disease," *Neurology and Therapy*, vol. 11, pp. 525–551, 2022. [Online]. Available: https://link.springer.com/article/10.1007/s40120-022-00335-x

[38] H. Tang, E. Yao, G. Tan, and X. Guo, "A fast and accurate 3d fine-tuning convolutional neural network for alzheimer's disease diagnosis," in *Artificial Intelligence*. Springer, 2018, pp. 115–126. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-13-2122-1_9

[39] M. Prince, A. Wimo, M. Guerchet, G.-C. Ali, Y.-T. Wu, and M. Prina, *World Alzheimer Report 2015: The global impact of dementia*. Alzheimer's Disease International, 2015.

[40] B. Dubois, H. Hampel, H. H. Feldman, P. Scheltens, P. Aisen, and S. Andrieu, "Preclinical alzheimer's disease: Definition, natural history, and diagnostic criteria," *Alzheimer's & Dementia*, vol. 12, no. 3, pp. 292–323, 2016.

[41] R. C. Petersen, B. Caracciolo, C. Brayne, S. Gauthier, V. Jelic, and L. Fratiglioni, "Mild cognitive impairment," *Nature Reviews Disease Primers*, vol. 1, p. 15003, 2014.

[42] D. J. Selkoe, "Alzheimer's disease: genes, proteins, and therapy," *Physiological Reviews*, vol. 81, no. 2, pp. 741–766, 2001.

[43] D. J. Selkoe and J. Hardy, "The amyloid hypothesis of alzheimer's disease at 25 years," *EMBO Molecular Medicine*, vol. 8, no. 6, pp. 595–608, 2016.

[44] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, R. C. Mohs, J. C. Morris, M. N. Rossor, P. Scheltens, M. C. Carrillo, B. Thies, S. Weintraub, and M. H. Phelps, "The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging–alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & Dementia*, vol. 7, no. 3, pp. 263–269, 2011. [Online]. Available: https://alz-journals.onlinelibrary.wiley.com/doi/10.1016/j.jalz.2011.03.005

[45] Alzheimer's Association. (2023) Etapas y síntomas de la enfermedad de alzheimer. [Online]. Available: https://www.alz.org/es-mx/alzheimer-demencia/etapas

[46] H. Braak and E. Braak, "Neuropathological stageing of alzheimer-related changes," *Acta Neuropathologica*, vol. 82, no. 4, pp. 239–259, 1991.

[47] B. Dubois, H. H. Feldman, C. Jacova, H. Hampel, J. L. Molinuevo, K. Blennow, S. T. DeKosky, S. Gauthier, D. Selkoe, R. Bateman *et al.*, "Advancing research diagnostic criteria for alzheimer's disease: the iwg-2 criteria," *The Lancet Neurology*, vol. 13, no. 6, pp. 614–629, 2014.

[48] E. M. Haacke, R. W. Brown, M. R. Thompson, and R. Venkatesan, *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. New York: Wiley-Liss, 1999.

[49] R. A. Pooley, "Fundamental physics of mr imaging," *Radiographics*, vol. 25, no. 4, pp. 1087–1099, 2005.

[50] J. P. Hornak, *The Basics of MRI*. Rochester: Interactive Learning Software, 1996.

[51] R. L. Wahl, J. M. Herman, and E. Ford, "The promise and pitfalls of positron emission tomography and molecular imaging," *Journal of Clinical Oncology*, vol. 26, no. 10, pp. 1734–1740, 2008.

[52] K. Herholz and K. Ebmeier, "Clinical amyloid imaging in alzheimer's disease," *The Lancet Neurology*, vol. 10, no. 7, pp. 667–670, 2011.

[53] L. Cai, S. Lu, and V. W. Pike, "Chemistry with [18f]fluoride ion," *European Journal of Organic Chemistry*, vol. 2008, no. 17, pp. 2853–2873, 2008.

[54] S. R. Cherry, J. A. Sorenson, and M. E. Phelps, *Physics in Nuclear Medicine*, 4th ed. Philadelphia: Elsevier Health Sciences, 2012.

[55] M. E. Phelps, "Positron emission tomography provides molecular imaging of biological processes," *Proceedings of the National Academy of Sciences*, vol. 97, no. 16, pp. 9226–9233, 2000.

[56] V. L. Villemagne, V. Doré, S. C. Burnham, C. L. Masters, and C. C. Rowe, "Imaging tau and amyloid-β proteinopathies in alzheimer disease and other conditions," *Nature Reviews Neurology*, vol. 14, no. 4, pp. 225–236, 2018.

[57] M. E. Phelps, S.-C. Huang, E. J. Hoffman, C. Selin, L. Sokoloff, and D. E. Kuhl, "Tomographic measurement of local cerebral glucose metabolic rate in humans with (f-18) 2-fluoro-2-deoxy-d-glucose: validation of method," *Annals of Neurology*, vol. 6, no. 5, pp. 371–388, 1979.

[58] L. Sokoloff, M. Reivich, C. Kennedy, M. Des Rosiers, C. Patlak, K. Pettigrew, O. Sakurada, and M. Shinohara, "The [14c]deoxyglucose method for the measurement of local cerebral glucose utilization: theory, procedure, and normal values in the conscious and anesthetized albino rat," *Journal of Neurochemistry*, vol. 28, no. 5, pp. 897–916, 1977.

[59] L. Mosconi, "Brain glucose metabolism in the early and specific diagnosis of alzheimer's disease: Fdg-pet studies in mci and ad," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 32, no. 4, pp. 486–510, 2005.

[60] W. E. Klunk, H. Engler, A. Nordberg, Y. Wang, G. Blomqvist, D. P. Holt, M. Bergström, I. Savitcheva, G.-f. Huang, S. Estrada *et al.*, "Imaging brain amyloid in alzheimer's disease with pittsburgh compound-b," *Annals of Neurology*, vol. 55, no. 3, pp. 306–319, 2004.

[61] W. J. Jansen, R. Ossenkoppele, D. L. Knol, B. M. Tijms, P. Scheltens, F. R. Verhey, P. J. Visser, P. Aalten, D. Aarsland, D. Alcolea *et al.*, "Prevalence of cerebral amyloid pathology in persons without dementia: a meta-analysis," *JAMA*, vol. 313, no. 19, pp. 1924–1938, 2015.

[62] C. M. Clark, M. J. Pontecorvo, T. G. Beach, B. J. Bedell, R. E. Coleman, P. M. Doraiswamy, A. S. Fleisher, E. M. Reiman, M. N. Sabbagh, C. H. Sadowsky *et al.*, "Cerebral pet with florbetapir compared with neuropathology at autopsy for detection of neuritic amyloid-β plaques: a prospective cohort study," *The Lancet Neurology*, vol. 11, no. 8, pp. 669–678, 2012.

[63] M. F. Folstein, S. E. Folstein, and P. R. McHugh, "Mini-mental state: a practical method for grading the cognitive state of patients for the clinician," *Journal of Psychiatric Research*, vol. 12, no. 3, pp. 189–198, 1975.

[64] T. N. Tombaugh and N. J. McIntyre, "The mini-mental state examination: a comprehensive review," *Journal of the American Geriatrics Society*, vol. 40, no. 9, pp. 922–935, 1992.

[65] M. W. Weiner, D. P. Veitch, P. S. Aisen, L. A. Beckett, N. J. Cairns, R. C. Green, D. Harvey, C. R. Jack Jr, W. Jagust, J. C. Morris *et al.*, "The alzheimer's disease neuroimaging initiative: a review of papers published since its inception," *Alzheimer's & Dementia*, vol. 11, no. 6, pp. e1–e120, 2015.

[66] R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, C. R. Jack Jr, W. J. Jagust, L. M. Shaw, A. W. Toga *et al.*, "Alzheimer's disease neuroimaging initiative (adni): clinical characterization," *Neurology*, vol. 74, no. 3, pp. 201–209, 2010.

[67] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2010.

[68] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, pp. 24–29, 2019.

[69] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[70] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[71] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[72] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[73] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems*, 2014.

[74] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[75] Y. Duan, X. Liu, Y. Liu, and Y. Zhao, "Hybrid deep learning models for medical image classification: A review," *IEEE Access*, vol. 9, pp. 103 512–103 531, 2021.

[76] M. Islam, Y. Xiaohui, A. Burry, and T. Gedeon, "Multimodal fusion techniques for biomedical diagnosis: A review," *Information Fusion*, vol. 76, pp. 56–78, 2022.

[77] Y. Gao, Q. Ruan, and Q. Wu, "Hybrid model combining deep cnn and classical classifiers for alzheimer's disease diagnosis using pet images," in *Proceedings of the 2020 International Conference on Medical Imaging and Computer-Aided Diagnosis (MICAD)*, 2020, pp. 25–30.

[78] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[79] A. W. Kushniruk and V. L. Patel, "Cognitive and usability engineering methods for the evaluation of clinical information systems," *Journal of Biomedical Informatics*, vol. 43, no. 5, pp. 739–750, 2010.

[80] *ISO 9241-210:2010 - Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems*, International Organization for Standardization, 2010.

[81] Q. Zhou, H. Song, and Q. Wang, "Design and implementation of a medical image classification and visualization system using deep learning and dash," *IEEE Access*, vol. 9, pp. 88 894–88 903, 2021.

[82] A. Rogeau, F. Hives, C. Bordier, H. Lahousse, V. Roca, T. Lebouvier, F. Pasquier, D. Huglo, F. Semah, and R. Lopes, "A 3d convolutional neural network to classify subjects as alzheimer's disease, frontotemporal dementia or healthy controls using brain 18f-fdg pet," *NeuroImage*, vol. 288, p. 120530, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811924000259

[83] M. Liu, D. Cheng, K. Wang, and Y. Wang, "Multi-modality cascaded convolutional neural networks for alzheimer's disease diagnosis," *Neuroinformatics*, vol. 18, no. 2, pp. 295–308, 2020.

[84] C. R. Jack Jr, D. S. Knopman, W. J. Jagust, L. M. Shaw, P. S. Aisen, M. W. Weiner, R. C. Petersen, and J. Q. Trojanowski, "Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade," *The Lancet Neurology*, vol. 9, no. 1, pp. 119–128, 2010.

[85] V. L. Villemagne, S. Burnham, P. Bourgeat, B. Brown, K. A. Ellis, O. Salvado, C. Szoeke, S. L. Macaulay, R. Martins, P. Maruff *et al.*, "Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic alzheimer's disease: a prospective cohort study," *The Lancet Neurology*, vol. 12, no. 4, pp. 357–367, 2013.

[86] R. Ossenkoppele, D. R. Schonhaut, M. Schöll, S. N. Lockhart, N. Ayakta, S. L. Baker, J. P. O'Neil, M. Janabi, A. Lazaris, A. Cantwell *et al.*, "Tau pet patterns mirror clinical and neuroanatomical variability in alzheimer's disease," *Brain*, vol. 139, no. 5, pp. 1551–1567, 2016.

[87] M. Schöll, S. N. Lockhart, D. R. Schonhaut, J. P. O'Neil, M. Janabi, R. Ossenkoppele, S. L. Baker, J. W. Vogel, J. Faria, H. D. Schwimmer *et al.*, "Pet imaging of tau deposition in the aging human brain," *Neuron*, vol. 89, no. 5, pp. 971–982, 2016.

[88] N. I. Bohnen, D. S. Djang, K. Herholz, Y. Anzai, and S. Minoshima, "Effectiveness and safety of 18f-fdg pet in the evaluation of dementia: a review of the recent literature," *Journal of Nuclear Medicine*, vol. 53, no. 1, pp. 59–71, 2012.

[89] X. Li, P. S. Morgan, J. Ashburner, J. Smith, and C. Rorden, "The first step for neuroimaging data analysis: Dicom to nifti conversion," *Journal of Neuroscience Methods*, vol. 264, pp. 47–56, 2016.

[90] D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellyn, and W. Eubank, "Pet-ct image registration in the chest using free-form deformations," *IEEE transactions on medical imaging*, vol. 22, no. 1, pp. 120–128, 2003.

[91] C.-C. Liu, T. Kanekiyo, H. Xu, and G. Bu, "Apolipoprotein e and alzheimer disease: risk, mechanisms and therapy," *Nature Reviews Neurology*, vol. 9, no. 2, pp. 106–118, 2013.

[92] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*, 3rd ed. John Wiley & Sons, 2019, vol. 793.

[93] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, no. 2, pp. 207–244, 2009.

[94] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 6, no. 4, pp. 1–21, 2012.

[95] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler *et al.*, "Api design for machine learning software: experiences from the scikit-learn project," *arXiv preprint arXiv:1309.0238*, 2013.

[96] S. Raschka, *Model evaluation, model selection, and algorithm selection in machine learning*, 2018.

[97] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2. Montreal, Canada, 1995, pp. 1137–1145.

[98] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[99] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[100] S. Chen, K. Ma, and Y. Zheng, "Med3d: Transfer learning for 3d medical image analysis," *arXiv preprint arXiv:1904.00625*, 2019.

[101] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[102] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018.

[103] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[104] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[105] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[106] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[107] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, "Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, 2018.

[108] K. P. Murphy, "Naive bayes classifiers," Vancouver, 2006. [Online]. Available: https://www.cs.ubc.ca/~murphyk/Teaching/CS340-Fall06/lectures/naiveBayes.pdf

[109] D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*, ser. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press, 2001.

[110] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[111] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, 2nd ed. Springer, 2009, vol. 2.

[112] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[113] A. Abid, A. Abdalla, A. Abid, D. Khan, A. Alfozan, and J. Zou, "Gradio: Hassle-free sharing and testing of ml models in the wild," *arXiv preprint arXiv:1906.02569*, 2019.

[114] Plotly Technologies Inc., "Collaborative data science," Montreal, QC, 2015. [Online]. Available: https://plot.ly

[115] Python Software Foundation, "Python language reference, version 3.x," 2023, accessed: 2024. [Online]. Available: https://www.python.org

[116] C. R. Harris, K. J. Millman, S. J. Van Der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith *et al.*, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[117] W. McKinney *et al.*, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, vol. 445, no. 1. Austin, TX, 2010, pp. 51–56.

[118] openpyxl Development Team, "openpyxl - a python library to read/write excel 2010 xlsx/xlsm files," 2023, version 3.0.x. [Online]. Available: https://openpyxl.readthedocs.io

[119] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[120] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.

[121] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[122] M. L. Waskom, "seaborn: statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.

[123] Python Software Foundation, "pickle — python object serialization," 2023, python 3.x documentation. [Online]. Available: https://docs.python.org/3/library/pickle.html

[124] joblib Development Team, "Joblib: running python functions as pipeline jobs," 2023, version 1.2.x. [Online]. Available: https://joblib.readthedocs.io

[125] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016, pp. 265–283.

[126] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026–8037.

[127] M. Brett, C. J. Markiewicz, M. Hanke, M.-A. Côté, B. Cipollini, P. McCarthy, D. Jarecka, C. P. Cheng, Y. O. Halchenko, M. Cottaar *et al.*, "Nibabel: Access a cacophony of neuro-imaging file formats," *Zenodo*, 2020.

[128] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright *et al.*, "Scipy 1.0: fundamental algorithms for scientific computing in python," *Nature methods*, vol. 17, no. 3, pp. 261–272, 2020.

[129] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek, "The design of simpleitk," *Frontiers in neuroinformatics*, vol. 7, p. 45, 2013.

[130] ReportLab Inc., "Reportlab: Pdf generation toolkit," *ReportLab Documentation*, 2020. [Online]. Available: https://www.reportlab.com/

[131] P. S. Foundation, "os miscellaneous operating system interfaces," https://docs.python.org/3/library/os.html, 2023, python 3.x documentation.

[132] P. Software, "json json encoder and decoder," https://docs.python.org/3/library/json.html, 2023, python 3.x documentation.

[133] Python Software Foundation, "warnings — warning control," 2023, python 3.x documentation. [Online]. Available: https://docs.python.org/3/library/warnings.html

[134] P. Software, "datetime basic date and time types," https://docs.python.org/3/library/datetime.html, 2023, python 3.x documentation.