Universitat Oberta de Catalunya (UOC)

Máster Universitario en Ciencia de Datos (*Data Science*)

TRABAJO FINAL DE MÁSTER

Área: *Sustainable computing*

Hacia un Diagnóstico Computacional de la Depresión: Un Enfoque Basado en Deep Learning Híbrido para el Análisis de Resonancias Magnéticas

---

Autor: Alejandro Francisco Chivite Bermúdez

Tutor: Raúl Parada Medina

Profesor: Susana Acedo Nadal

Pamplona, 28 de mayo de 2025

Créditos/Copyright

FICHA DEL TRABAJO FINAL

Título del trabajo:	Hacia un Diagnóstico Computacional de la Depresión: Un Enfoque Basado en Deep Learning Híbrido para el Análisis de Resonancias Magnéticas
Nombre del autor:	Alejandro Francisco Chivite Bermudez
Nombre del colaborador/a docente:	Raúl Parada Medina
Nombre del PRA:	Susana Acedo Nadal
Fecha de entrega (mm/aaaa):	28 de mayo de 2025
Titulación o programa:	Máster Universitario en Ciencia de Datos (*Data Science*)
Área del Trabajo Final:	*Sustainable Computing*
Idioma del trabajo:	Español
Palabras clave	*Deep Learning*, Neuroimagen, Depresión

Agradecimientos

Quisiera agradecer a mi tutor Raúl Parada Medina por su constante apoyo y disposición a resolver mis numerosas dudas. También, por haberme brindado la oportunidad de desarrollar este proyecto de elección propia, el cual me ha permitido integrar mis conocimientos adquiridos durante mi formación en Psicología junto con los obtenidos en este Máster de Ciencia de Datos. Asimismo, agradezco profundamente a los autores Bezmaternykh et al., (2021) por hacer sus datos accesibles de forma abierta, lo que no solo ha hecho posible el desarrollo del proyecto, sino que también contribuye al avance del diagnóstico y la comprensión de la depresión. Por último, quiero agradecer a mi tía Leila Chivite, cuya inspiración fue decisiva para emprender este camino académico.

Resumen

La depresión es el trastorno mental más común y su diagnóstico se basa en evaluaciones clínicas que presentan, en ocasiones, un elevado porcentaje de error. En este contexto, la neuroimagen, junto con la aplicación de modelos de *deep learning*, emerge como una herramienta prometedora. Sin embargo, su aplicación sigue siendo limitada debido a la falta de modelos robustos e interpretables.

El presente proyecto propone el desarrollo de un modelo de *deep learning* híbrido que combine redes neuronales convolucionales y recurrentes con el objetivo de clasificar correctamente imágenes de resonancia magnética funcional en reposo de personas con y sin diagnóstico de depresión. Además, para mejorar el rendimiento y la interpretabilidad del modelo, se han implementado diferentes estrategias del estado del arte como *transfer learning*, *data augmentation*, técnicas de *explainable artificial intelligence* y técnicas de manejo de etiquetas erróneas.

Los resultados muestran un alto rendimiento general de la arquitectura híbrida propuesta junto con la aplicación de *transfer learning*. El mejor modelo (basado en MobileNetV2) alcanzó una sensibilidad del 81.25 % y un AUC de 0.90. Por otro lado, ni el *data augmentation* ni las técnicas de manejo de etiquetas erróneas lograron mejorar estos resultados. Sin embargo, el uso de *integrated gradients* como técnica XAI sí permitió identificar regiones espaciales y temporales relevantes.

En conclusión, las arquitecturas y técnicas implementadas muestran un alto potencial como herramienta complementaria para el diagnóstico de la depresión.

Palabras clave: Depresión, Diagnóstico Erróneo, Neuroimagen, Aprendizaje Profundo, Redes Neuronales Convolucionales, Redes Neuronales Recurrentes, Transferencia de Aprendizaje, Inteligencia Artificial Explicable.

Abstract

Depression is the most common mental disorder and its diagnosis is based on clinical assessments that sometimes have a high error rate. In this context, neuroimaging, together with the application of deep learning models, emerges as a promising tool. However, its application remains limited due to the lack of robust and interpretable models.

The present project proposes the development of a hybrid deep learning model combining convolutional and recurrent neural networks with the aim of correctly classifying resting functional magnetic resonance images of people with and without a diagnosis of depression. Furthermore, in order to improve the performance and interpretability of the model, different state-of-the-art strategies such as transfer learning, data augmentation, explainable artificial intelligence techniques and noisy labels handling techniques have been implemented.

The results show an overall high performance of the proposed hybrid architecture together with the application of transfer learning. The best model (based on MobileNetV2) achieved a sensitivity of 81.25 % and an AUC of 0.90. On the other hand, neither data augmentation nor noisy labels handling techniques were able to improve these results. However, the use of integrated gradients as an XAI technique did identify relevant spatial and temporal regions.

In conclusion, the implemented architectures and techniques show a high potential as a complementary tool for the diagnosis of depression.

Keywords: Depression, Misdiagnosis, Neuroimaging, Deep Learning, Convolutional Neural Networks, Recurrent Neural Networks, Transfer Learning, Explainable Artificial Intelligence.

Índice general

Lista de figuras

Lista de tablas

Capítulo 1

Introducción

Este capítulo presenta el contexto y la justificación del proyecto. Además, se detalla la metodología empleada, incluyendo la estrategia de investigación, los datos utilizados y el proceso de implementación. También, se abordan las implicaciones éticas, sociales y de sostenibilidad del proyecto. Por último, se describe su planificación temporal.

1.1. Contexto y justificación

En esta sección se ofrece una visión general de los principales aspectos que justifican el desarrollo del proyecto. Se analiza el impacto creciente de la depresión a nivel global, el papel de la neuroimagen como herramienta clave para explorar el cerebro, y el desarrollo de modelos computacionales para la detección de trastornos mentales. Asimismo, se abordan las limitaciones actuales en la práctica diagnóstica y se destaca el potencial de las tecnologías emergentes para transformar este campo.

1.1.1. Un trastorno en aumento y un desafío global

La depresión es el trastorno psicológico más común, afecta a cerca de 350 millones de personas en todo el mundo (Lim et al. 2018) y su incidencia está en aumento (Moreno-Agostino et al. 2021). Esta condición se caracteriza por una tristeza patológica que provoca alteraciones físicas y cognitivas en la persona, afectando así a aspectos fundamentales como el desarrollo funcional, el lenguaje y las relaciones sociales. Según la Organización Mundial de la Salud (World Health Organization 2017), la depresión es actualmente la segunda mayor causa de discapacidad en el mundo y se espera que para 2030 se convierta en la primera. Por esta razón, es imperativo buscar soluciones innovadoras para reducir el impacto de este trastorno y mejorar la calidad de vida de cientos de miles de personas.

1.1.2. Una herramienta clave para comprender el cerebro

La neuroimagen, es decir, la obtención de imágenes cerebrales mediante diversas técnicas para entender la estructura y el funcionamiento cerebral, se ha establecido como el principal método para el diagnóstico y el tratamiento de distintas enfermedades y condiciones, especialmente las neurodegenerativas (Chouliaras & O'Brien 2023). Sin embargo, su uso para el diagnóstico y tratamiento de trastornos psicológicos es escaso. Esta falta de aplicación se debe a que todavía existen dudas respecto a la etiología y la fisiopatología de este tipo de trastornos (Prvulovic & Hampel 2010). Es decir, todavía no se comprende plenamente cómo surgen estas condiciones y cómo afectan al cerebro. De hecho, la Asociación Americana de Psicología (First et al. 2018) desincentiva actualmente su uso para el diagnóstico y el tratamiento de trastornos psicológicos.

A pesar de ello, recientemente, numerosos estudios han logrado grandes avances en este campo al descubrir diferencias en distintas regiones y conexiones neuronales en personas con y sin diagnóstico de depresión (Henderson et al. 2020). Por ejemplo, como comentan Dunlop & Mayberg (2017), recientes estudios han identificado volúmenes reducidos en el hipocampo, una menor activación en la corteza prefrontal dorsolateral y una mayor activación en áreas como la amígdala y la corteza cingulada subcallosa en personas que padecen este trastorno.

1.1.3. Modelos para la detección de trastornos mentales

Más recientemente, numerosos estudios han empezado a desarrollar modelos de *machine learning* (ML) entrenados con conjuntos de datos de neuroimagen (Smucny, Shi & Davidson 2022), debido a su gran capacidad para procesar conjuntos de datos multidimensionales y realizar predicciones precisas. El objetivo de estas investigaciones es obtener modelos capaces de identificar diferentes condiciones mediante el análisis de los resultados de las neuroimágenes obtenidas de un paciente. Por ejemplo, el metaanálisis realizado por Quaak et al. (2021) identificó que los modelos de clasificación del trastorno del espectro autista obtenían resultados prometedores, con precisiones de hasta el 94 %.

A pesar de estos buenos resultados, es imprescindible tener en cuenta las limitaciones de estos estudios, entre las que se incluyen datos obtenidos de diferentes centros, muestras reducidas y heterogéneas, y la falta de información sobre los criterios que utilizan los modelos para la clasificación (Eitel et al. 2021).

1.1.4. Errores de diagnóstico en la salud mental

En el campo de la salud, existe un cierto porcentaje de diagnósticos erróneos. Según Schiff et al. (2009), el diagnóstico erróneo es cualquier equivocación o fallo en el proceso de diagnósti-

co que conduce a un diagnóstico erróneo, omitido o tardío. La prevalencia de este tipo de diagnósticos depende de diferentes factores, como el tipo de enfermedad, el perfil de la persona atendida, o el entorno médico, entre otros (Ball et al. 2015). En términos generales, diferentes estudios estiman que el porcentaje se sitúa entre el 5 % y el 15 % de los casos (Lorenzo et al. 2021). Esta situación conlleva graves consecuencias, como el aumento del coste de los servicios sanitarios, la aplicación de tratamientos inadecuados o innecesarios y la reducción de la calidad de vida de las personas, entre otras.

En el campo de la salud mental, la tasa de prevalencia de los diagnósticos erróneos es aún mayor, suponiendo alrededor de un tercio de los casos, aunque para algunos trastornos este porcentaje es aún mayor (Ayano et al. 2021). Según Vermani et al. (2011), las tasas de diagnóstico erróneo en atención primaria en el caso del trastorno depresivo mayor alcanzan el 65,9 %.

Debido a las graves consecuencias que conlleva un alto porcentaje de diagnósticos erróneos, es primordial desarrollar mecanismos que reduzcan considerablemente estas tasas. Por lo tanto, la creación de un modelo preciso, capaz de identificar la depresión mediante la examinación de imágenes cerebrales podría ser de gran utilidad.

1.1.5. Tecnologías emergentes en neuroimagen

Actualmente, la neuroimagen tiene un alto coste, tanto en su adquisición como en su aplicación. Por un lado, la inversión inicial es muy elevada. Dependiendo del tipo de escáner, su precio puede oscilar entre varios cientos de miles y millones de euros (Sarracanie et al. 2015). Por otro lado, su utilización también es costosa debido a la gran cantidad de energía que estos aparatos consumen y al equipo necesario para hacerlos funcionar (Signorile et al. 2024).

Afortunadamente, en los ultimos años diversos proyectos en el mundo están trabajando en la creación de nuevos equipos de neuroimagen que, a pesar de tener una precisión menor, tienen un coste y un tamaño significativamente más bajos (Wald et al. 2020). Esto permitirá que países y regiones que no tenían acceso a esta tecnología puedan finalmente acceder a ella.

1.2. Motivación

El motivo del desarrollo de este proyecto es la confluencia de los temas tratados anteriormente. En primer lugar, la depresión tiene una gran incidencia en la sociedad, problemática que se agrava además debido al alto porcentaje de diagnósticos erróneos. En segundo lugar, constantemente están surgiendo grandes avances en el campo de la neuroimagen y el ML, lo que resulta en modelos cada vez más precisos. En tercer lugar, también están surgiendo nuevas alternativas de neuroimagen más baratas y accesibles.

Por lo tanto, la creación de un modelo de ML preciso podría ser de gran utilidad para la sociedad, ya que podría ayudar a reducir el número de diagnósticos erróneos. El desarrollo de este modelo, junto con la reducción del coste de la tecnología de neuroimagen, permitiría que las personas que viven en las zonas con menos recursos del planeta también pudieran acceder a un diagnóstico psicológico preciso y fiable.

1.2.1. Objetivo

El objetivo del presente proyecto es desarrollar un modelo de ML capaz de identificar si una paersona padece depresión mediante el procesamiento de las neuroimágenes obtenidas de este. Los datos de neuroimagen con los que se entrenará el modelo serán datos de resonancia magnética (RM). Esta modalidad de neuroimagen es la más utilizada en el campo de los trastornos psicológicos y ofrece la mayor precisión (Najafpour et al. 2021). Más específicamente, se hará uso de datos de *resonancia magnética funcional* (RMf), ya que como se ha comentado anteriormente, recientes estudios sugieren que los cerebros de personas con depresión pueden presentar alteraciones tanto en su estructura como en su funcionamiento, y los datos de RMf proveen de información tanto espacial como temporal.

El modelo de ML que se desarrollará será un modelo de *deep learning* (DL), ya que recientes investigaciones sugieren que este tipo de modelos obtienen mejores resultados que otros modelos de ML tradicional como las máquinas de soporte vectorial o los árboles de decisión (Quaak et al. 2021).

Además, siguiendo las pautas y recomendaciones del estado del arte en este campo (Smucny, Shi & Davidson 2022), se desarrollará un modelo de DL que supere las limitaciones típicas de este tipo de modelos, entre las que se incluyen: mejorar la capacidad de generalización en conjuntos de datos limitados mediante *transfer learning*, tratar con muestras reducidas mediante *data augmentation*, y abordar la falta de comprensión de las características que utiliza el modelo para tomar sus decisiones mediante *explainable artificial intelligence* (XAI).

Las principales medidas de evaluación de modelos aplicados a la salud son la sensibilidad y la especificidad. Por un lado, la sensibilidad es especialmente importante en los modelos aplicados al diagnóstico de enfermedades, ya que el objetivo es evitar los falsos negativos. Es decir, se pretende evitar que una persona afectada no reciba su tratamiento. Por otro lado, la especificidad es especialmente importante en modelos que recomienden tratamientos específicos, ya que el objetivo es evitar los falsos positivos. Es decir, se pretende evitar que un paciente sano reciba un tratamiento con graves efectos secundarios. Por lo tanto, dado que el modelo tendrá aplicaciones en el campo de la salud, y como cuyo objetivo es el diagnóstico, la medida de evaluación principal será la sensibilidad.

Por último, dado que el campo de la psicología presenta, en ocasiones, un número considera-

ble de diagnósticos clínicos imprecisos, se aplicarán también diferentes técnicas de tratamiento del ruido en las etiquetas (en inglés, *noisy labels*). Estas técnicas permiten mitigar el impacto negativo que puedan tener los errores de diagnóstico en el proceso de aprendizaje del modelo.

En conclusión, se desarrollará un modelo de DL con datos de neuroimagen, aplicando técnicas del estado del arte en el campo, con el objetivo de clasificar de forma precisa a personas con y sin diagnóstico de depresión.

1.3. Metodología

El presente proyecto sigue una estrategia de diseño y construcción. Más concretamente, se desarrollará un modelo híbrido de redes neuronales convolucionales (CNN) y recurrentes (RNN) para clasificar imágenes de RMf de personas con y sin diagnóstico de depresión. La elección de un modelo híbrido se debe a que los datos de RMf contienen información tanto espacial como temporal. Por un lado, la información espacial representa la activación cerebral en diferentes intensidades, para lo cual las CNN son ideales. Por el otro lado, la información temporal representa cómo cambia la actividad cerebral con el tiempo, para lo cual las RNN son la mejor opción. Dado que, como se ha comentado anteriormente, estudios previos sugieren que los cerebros de personas con diagnóstico de depresión presentan diferencias tanto estructurales como funcionales, la mejor opción es combinar estos modelos para obtener un modelo más robusto que obtenga una mejor clasificación y que capture la verdadera naturaleza de esta condición.

En primer lugar, se implementan una serie de tareas de preprocesamiento de datos con el objetivo de prepararlos para los modelos de DL posteriores. A continuación, se desarrollan una serie de experimentos para examinar el impacto que tienen las técnicas mencionadas en el aumento de la robustez y precisión de los modelos de DL. Todos los modelos se evalúan según las métricas estándar, en particular la sensibilidad. Por último, se implementan técnicas XAI para identificar las regiones cerebrales clave en la clasificación.

1.4. Competencia de compromiso ético y global (CCEG) y objetivos de desarrollo sostenible (ODS)

El proyecto tiene en consideración las tres dimensiones clave de la CCEG (sostenibilidad, responsabilidad social y diversidad) mediante la alineación con varios ODS de la Agenda 2030 de la ONU (Lee et al. 2016).

1.4.1. Sostenibilidad

Un aspecto importante es a tener en cuenta es el alto consumo energético asociado a los escáneres de resonancia magnética, ya que una única resonancia equivale al consumo energético de un mes de televisión (Signorile et al. 2024). Esto implica que el conjunto de datos utilizado para este proyecto tiene una huella ecológica significativa.

Sin embargo, al tratarse de datos de acceso abierto, se fomenta la reutilización de recursos, lo que optimiza el uso de los datos ya generados sin necesidad de realizar nuevas exploraciones. Además, se han tenido en consideración los recientes avances en el desarrollo de equipos de neuroimagen más eficientes desde el punto de vista energético, con el objetivo de que el modelo desarrollado pueda aplicarse a datos obtenidos de estos nuevos dispositivos en un futuro. Por lo tanto, el proyecto se alinea con el ODS 9 (Industria, innovación e infraestructura) y el ODS 12 (Producción y consumo responsables) ya que se propone un uso eficiente de recursos tecnológicos y datos abiertos.

1.4.2. Comportamiento ético y responsabilidad social

También se ha tenido en cuenta el impacto en los profesionales de la salud mental. El propósito del presente proyecto no es reemplazar la labor de los profesionales que realizan los diagnósticos, sino proporcionar apoyo en la toma de decisiones clínicas. Igualmente, el tratamiento psicológico seguirá estando siempre bajo la responsabilidad de los especialistas.

Además, se han tenido en cuenta las implicaciones éticas derivadas del uso de la inteligencia artificial en el ámbito de la salud. Por esta razón, se ha garantizado el desarrollo de un modelo transparente y comprensible. Este aspecto conlleva que el proyecto también se alinee con el ODS 3 (Salud y bienestar) al proponer herramientas innovadoras para mejorar el diagnóstico en salud mental. También, se alinea con el ODS 16 (Paz, justicia e instituciones sólidas) al fomentar prácticas diagnósticas más transparentes y equitativas.

1.4.3. Diversidad y derechos humanos

La implementación de un modelo de diagnóstico preciso puede contribuir a la equidad en salud mental, ya que actualmente existen sesgos de género (Floyd 1997) y de etnicidad (Teplin et al. 2023) en el diagnóstico de trastornos mentales. Es decir, ciertos grupos poblacionales tienen una mayor probabilidad de recibir un diagnóstico erróneo, por lo que un modelo fiable y preciso podría mitigar estos sesgos.

Sin embargo, es importante señalar que el conjunto de datos utilizado en el presente proyecto cuenta con una mayor representación de personas pertenecientes al sexo biológico femenino. Dado que existen diferencias estructurales y funcionales en el cerebro asociadas al sexo biológico

(DeCasien et al. 2022), los resultados obtenidos deben interpretarse con cautela. Es posible que el modelo presente una mayor precisión en la detección de la depresión en personas con ciertas características. Entonces, el proyecto también se alinea con el ODS 5 (Igualdad de género) y el ODS 10 (Reducción de las desigualdades) al abordar los sesgos existentes en los diagnósticos clínicos.

En conclusión, aunque el proyecto presenta diferentes desafíos, el objetivo principal es desarrollar una herramienta que complemente el diagnóstico clínico, reduzca los sesgos y las desigualdades en la salud mental, y promueva el desarrollo de modelos con el menor impacto ecológico posible.

1.5. Planificación

En esta sección se presenta la planificación temporal del proyecto. Este se desarrollará en cinco fases principales, cada una compuesta por una serie de tareas clave. Además, al final de cada fase se requerirá la elaboración de distintos entregables, con el objetivo de documentar los avances obtenidos en cada fase.

- **M1 – Definición y planificación del trabajo final:** se establece la temática del proyecto y se justifica su relevancia. Además, se definen los objetivos principales y la planificación temporal del trabajo. En esta fase se deberá entregar una propuesta del proyecto.
- **M2 – Estado del arte:** se justifica con evidencia científica la línea de investigación escogida, se refinan los objetivos parciales del proyecto y se establece la metodología a desarrollar. En esta fase se entregará un documento que especifique los aspectos mencionados.
- **M3 – Diseño e implementación del trabajo:** se desarrolla e implementa el proyecto. En esta fase se entregará un documento que detalle los materiales, métodos y resultados obtenidos, así como los *notebooks* con la implementación.
- **M4 – Redacción de la documentación:** se elabora la memoria final del proyecto y una presentación audiovisual breve del mismo.
- **M5 – Defensa del proyecto:** se entrega la documentación del proyecto y se realiza una exposición oral del trabajo realizado ante el tribunal académico.

El siguiente diagrama de Gantt (Figura 1.1) se muestra la planificación temporal de los entregables mencionados, junto con las subtareas necesarias para su realización.

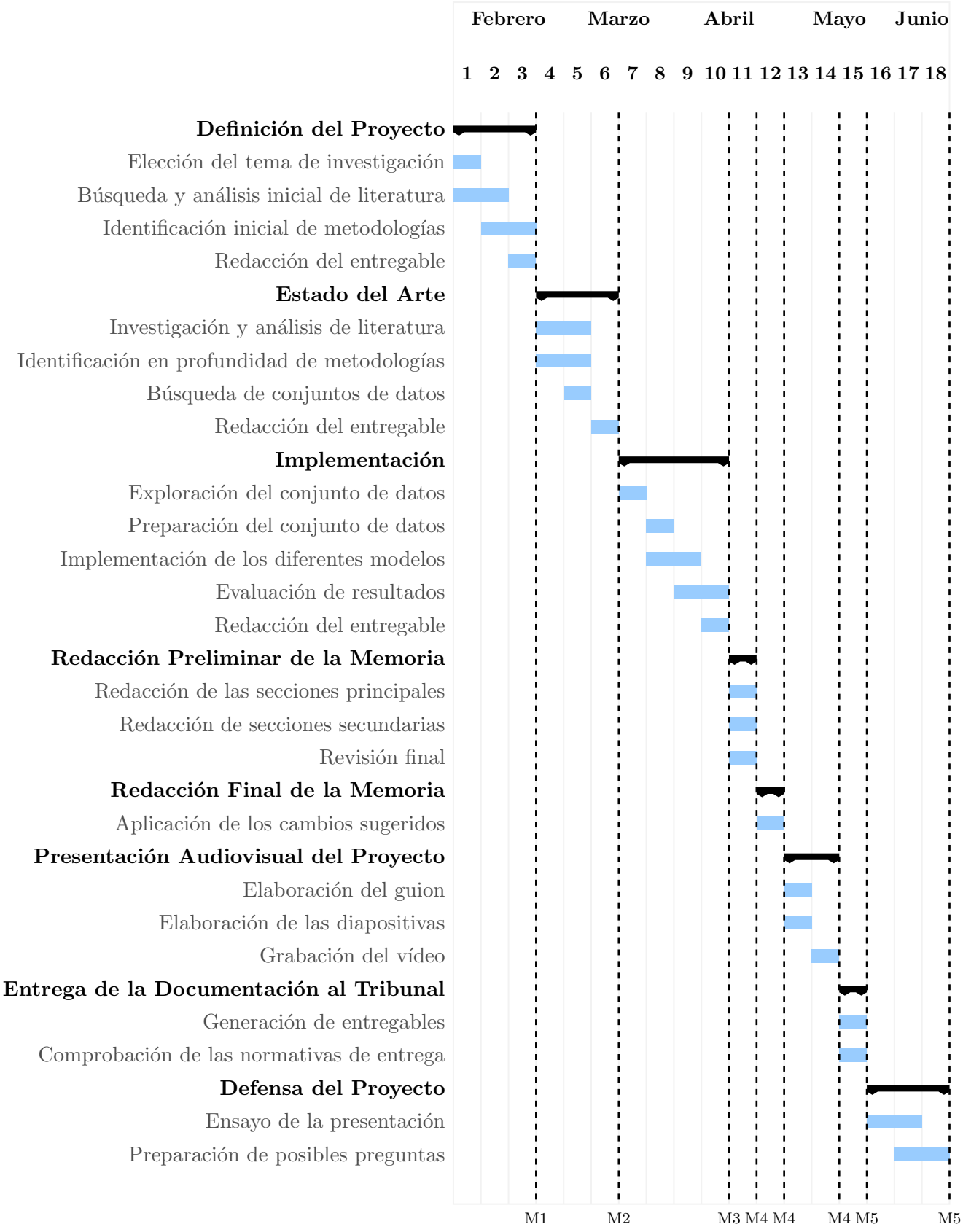


Figura 1.1: Planificación temporal del proyecto.

Capítulo 2

Estado del arte

En este capítulo se analizan los principales avances en el uso de neuroimagen y técnicas de ML aplicadas al estudio de la depresión. Se exploran las tecnologías de neuroimagen más utilizadas en psiquiatría en la actualidad, los últimos descubrimientos sobre los biomarcadores de la depresión y el papel de modelos como las CNN y las RNN en este campo. Además, se exponen técnicas actuales para mejorar el rendimiento de los modelos y se abordan los retos derivados del uso de *noisy labels*.

2.1. Neuroimagen en psiquiatría

La neuroimagen consiste en la obtención de imágenes cerebrales mediante diversas técnicas con el objetivo de estudiar su estructura y funcionamiento. En la psiquiatría, su principal finalidad es identificar los biomarcadores cerebrales asociados con los distintos trastornos psiquiátricos. Estos biomarcadores son características objetivas y medibles del cuerpo humano que permiten identificar procesos fisiológicos y patológicos (García-Gutiérrez et al. 2020). Por ejemplo, en el caso de la diabetes el nivel de glucosa en sangre constituye un biomarcador clínico. En el campo de la psiquiatría, los biomarcadores cerebrales pueden clasificarse en tres tipos: estructurales (volumen, forma, etc.), funcionales (activación) y de conectividad (relaciones entre regiones).

2.1.1. Retos actuales en el estudio de biomarcadores

Comprender cuáles biomarcadores están relacionados con los distintos trastornos psiquiátricos es esencial para su comprensión, prevención, diagnóstico, y tratamiento. Actualmente, el estudio de estos biomarcadores cerebrales está en auge, con más de 5000 artículos publicados en los últimos años (Nour et al. 2022). No obstante, las bases neuronales de la depresión siguen sin

estar completamente claras debido a varias razones. En primer lugar, los trastornos psiquiátricos son condiciones altamente complejas influenciadas por factores ambientales, socioculturales, psicológicos y genéticos (Nour et al. 2022). En segundo lugar, los estudios actuales presentan una serie de limitaciones metodológicas importantes entre las que se incluyen tamaños de efecto pequeños, movimientos del paciente durante el estudio, falta replicabilidad, enfoques puramente asociativos y tamaños muestrales reducidos (Kalin 2021). En tercer lugar, la comorbilidad (coexistencia de dos o más diagnósticos) es frecuente en psiquiatría (Henderson et al. 2020). Por ejemplo, la depresión presenta una elevada comorbilidad con la ansiedad (McElroy et al. 2018). Además, existen otras condiciones no psiquiátricas que pueden presentar síntomas que se confundan con trastornos psiquiátricos como la inflamación, las interacciones entre el intestino y el cerebro, o las lesiones cerebrales (Henderson et al. 2020).

2.1.2. Biomarcadores cerebrales de la depresión

A pesar de estos retos, recientemente numerosos estudios han logrado avances significativos en la identificación de biomarcadores tanto estructurales como funcionales asociados a la depresión. El metaanálisis de Gray et al. (2020) recopiló hallazgos consistentes entre distintos estudios, observando anomalías volumétricas en la corteza cingulada anterior subgenual, reducción del volumen del hipocampo izquierdo, activación anómala de la amígdala y alteraciones volumétricas en el putamen. De forma complementaria, Castanheira et al. (2019) también identificaron hallazgos similares, además de observar una mayor activación en la corteza cingulada anterior (CCA), la corteza prefrontal medial (CPPM) y la corteza cingulada posterior (CCP). Por su parte, el estudio multigrupo Schmaal et al. (2020) identificó, entre otros, un volumen reducido en el hipocampo y alteraciones en la activación en la CCA.

Estas regiones mencionadas están implicadas en algunas de las siguientes tres grandes redes cerebrales: la red de modo por defecto (RMD), la red de control cognitivo (RCC) y la red afectiva (RA), las cuales se encargan de la autorreflexión y la conciencia, así como de las tareas cognitivas que requieren atención y del procesamiento de las emociones respectivamente (Castanheira et al. 2019).

En la Tabla 2.1 se resumen estos hallazgos mencionados. Nótese que se incluyen únicamente aquellas estructuras y redes que presentan un mayor consenso en la literatura por su relación con la depresión, aunque no son las únicas implicadas.

Aunque numerosos estudios coinciden en la presencia de ciertas alteraciones cerebrales, también existen resultados divergentes entre investigaciones. Por este motivo, el uso de técnicas de DL representa una oportunidad para descubrir patrones latentes que no son fácilmente identificables mediante métodos estadísticos tradicionales.

Estructura	Anomalía
Corteza cingulada anterior	Hiperactivación y anomalías volumétricas
Hipocampo	Reducción volumétrica
Amígdala	Hiperactivación y reducción volumétrica
Putamen	Anomalías volumétricas
Corteza prefrontal medial	Hiperactivación
Red de modo por defecto	Alteraciones conectivas
Red de control cognitivo	Hipoactivación
Red afectiva	Hiperactivación

Tabla 2.1: Biomarcadores asociados con la depresión.

2.2. Técnicas de neuroimagen

En la actualidad existe un amplio abanico de técnicas de neuroimagen. Algunas de las más utilizadas se incluyen a continuación en orden cronológico:

- Electroencefalografía (EEG): detección de la actividad eléctrica cerebral mediante pequeños electrodos colocados sobre el cuero cabelludo (Yen et al. 2023).
- Tomografía computarizada (TC): obtención de imágenes transversales y tridimensionales del cerebro mediante rayos X (Ritt 2022).
- Tomografía de emisión de positrones (TEP): utiliza un agente radioactivo que el paciente ingiere, el cual se adhiere a la glucosa en sangre (Ramu & Haldorai 2024). Dado que el cerebro utiliza la glucosa como fuente de energía, el agente se acumula en las zonas con mayor actividad metabólica, permitiendo observar su distribución en el cerebro.
- Resonancia magnética (RM): utiliza campos magnéticos y ondas de radio para modificar la orientación de los protones de hidrógeno en las moléculas de agua en el cerebro. Al retornar a su estado original, estos protones emiten señales que permiten construir imágenes transversales de alta resolución (Yen et al. 2023).
- Magnetoencefalografía (MEG): medición de los campos magnéticos generados por la actividad eléctrica cerebral (Ramu & Haldorai 2024).
- Resonancia magnética funcional (RMf): medición de cambios en el flujo sanguíneo cerebral asociados a la actividad neuronal (Yen et al. 2023).

Como se puede observar, cada técnica de neuroimagen realiza distintas mediciones del cerebro, lo que se traduce en diferencias tanto en funcionales como operativas. Algunos de los

aspectos más relevantes a tener en cuenta incluyen el tipo de información a obtenida (estructural y/o funcional), el tipo de medida (directa o indirecta), la resolución espacial y temporal, y el aspecto invasivo de la técnica. También pueden considerarse otros aspectos como la duración del procedimiento, su coste, la portabilidad, la tolerancia del paciente o el impacto medioambiental (Najafpour et al. 2021, Signorile et al. 2024). En la Tabla 2.2 se resumen algunas de estas características clave de algunas de las técnicas mencionadas.

Técnica	**Tipo de información**	**Tipo de medida**	**Resolución espacial**	**Resolución temporal**	**Invasividad**
TC	Estructural	Directa	Alta	N/A	Invasiva (radiación ionizante)
TEP	Estructural/funcional	Indirecta	Baja	Muy baja (minutos)	No invasiva (radiofármaco)
EEG	Funcional	Directa	Baja	Alta (milisegundos)	No invasiva
MEG	Funcional	Directa	Alta	Alta (milisegundos)	No invasiva
RM	Estructural	Directa	Muy alta	N/A	No invasiva
RMf	Funcional	Indirecta	Alta	Baja (segundos)	No invasiva

Tabla 2.2: Comparativa de técnicas de neuroimagen.

Actualmente la técnica de neuroimagen más utilizada es la RMf. Esto se debe a que ofrece un equilibrio óptimo entre resolución espacial y temporal, está ampliamente disponible, tiene un coste moderado por sesión y no es invasiva (Bunge & Kahn 2009). Por estas razones, la mayoría de los estudios de neuroimagen aplicada a la psiquiatría también recurren a la RMf como técnica principal (Castanheira et al. 2019).

2.2.1. La resonancia magnética funcional

Más específicamente, la RMf permite estudiar la actividad cerebral de forma indirecta mediante la medición de cambios en el flujo sanguíneo. Cuando una región cerebral se activa, esta requiere de mayor energía y oxígeno, lo que resulta en un aumento del flujo sanguíneo en esa zona. A este aumento se la denomina respuesta hemodinámica (Glover 2011). De esta manera, la mayoría de los estudios de RMf se basan en el contraste *BOLD (Blood Oxygen Level Dependent)*, el cual detecta cambios en el contenido de oxígeno de la hemoglobina. La hemoglobina desoxigenada produce alteraciones magnéticas, las cuales son registradas por la resonancia magnética y permiten inferir que regiones están activadas (Glover 2011). Además, los estudios de RMf pueden realizarse mientras el paciente realiza ciertas tareas cognitivas, para observar que regiones se activan en relación con diferentes funciones, o mientras está en reposo, con el

objetivo de estudiar los patrones de conectividad funcional espontánea.

En conclusión, la RMf es una herramienta clave para el estudio de la estructura y el funcionamiento cerebral en personas con y sin un diagnóstico psiquiátrico como la depresión. Además, en los últimos años se están logrando avances significativos en el desarrollo de escáneres más potentes (Castanheira et al. 2019), lo cual permitirá obtener resultados más precisos en un futuro cercano.

2.3. *Machine learning* aplicado a la neuroimagen

El uso de técnicas de ML en neuroimagen comenzó a consolidarse en la década del 2010, impulsado principalmente por tres factores: el refinamiento de las diferentes técnicas de ML, el aumento de la potencia computacional y la aparición de grandes bases de datos de neuroimagen (Sánchez Fernández & Peters 2023).

El ML hace referencia a un conjunto de algoritmos que tienen la capacidad de aprender y mejorar de forma iterativa. Como señala Rajula et al. (2020), estos algoritmos presentan varias ventajas frente a los métodos estadísticos convencionales. En primer lugar, no requieren de hipótesis previas sobre la distribución de los datos, lo cual permite una mayor flexibilidad y capacidad de adaptación. Además, pueden manejar un gran número de variables y detectar interacciones complejas entre ellas. También permiten la integración de datos de distinta tipología.

Estas ventajas explican por qué los algoritmos de ML han ganado recientemente protagonismo frente a los métodos estadísticos tradicionales en este ámbito (Janssen et al. 2018).

2.3.1. Modelos convencionales de *machine learning*

Como destaca Zhao et al. (2023), entre un amplio abanico, los modelos de ML más utilizados en neuroimagen tradicionalmente han sido:

- *Random Forests (RF)*: un algoritmo de clasificación compuesto por numerosos arboles de decisión que hacen la función de clasificadores y que se entrenan en paralelo. El algoritmo integra los resultados de todos los árboles de decisión y asigna la categoría con más votos como resultado final.

- *Support vector machines (SVM)*: un algoritmo de clasificación y regresión que proyecta los datos de entrada en un espacio multidimensional y busca el hiperplano que maximice la separación entre clases mediante la utilización de funciones *kernel*.

Sin embargo, a pesar de su utilidad y de los grandes avances conseguidos en el campo de la neuroimagen mediante su uso (Janssen et al. 2018), estos algoritmos también presentan limitaciones. Según recientes estudios (Zhao et al. 2023, Sánchez Fernández & Peters 2023), estos algoritmos dependen de una selección manual de características, tienen dificultades para capturar relaciones no lineales complejas y requieren de una laboriosa preparación de datos, entre otros.

2.4. *Deep Learning* en neuroimagen

Con el objetivo de superar las limitaciones de los modelos de ML convencionales, en los últimos años numerosos estudios han optado por el uso modelos de DL.

El DL es una rama del ML que utiliza modelos de redes neuronales de múltiples capas, en los que cada capa procesa los datos con un nivel de abstracción mayor (Sánchez Fernández & Peters 2023).

Una red neuronal artificial (ANN, por sus siglas en inglés) es un algoritmo de ML que está compuesto por un conjunto de nodos que procesan y transforman datos. Cada nodo realiza una transformación simple de los datos y transmite el resultado al siguiente nodo. Estas estructuras se organizan en capas, de modo que cada capa transforma los datos recibidos de la capa anterior y envía su salida a la siguiente capa (Sánchez Fernández & Peters 2023). La estructura básica de estos modelos puede observarse en la Figura 2.1, donde se representan las tres partes principales de una ANN: la capa de entrada, las capas ocultas y la capa de salida, así como el flujo de información entre los nodos a través de conexiones dirigidas.

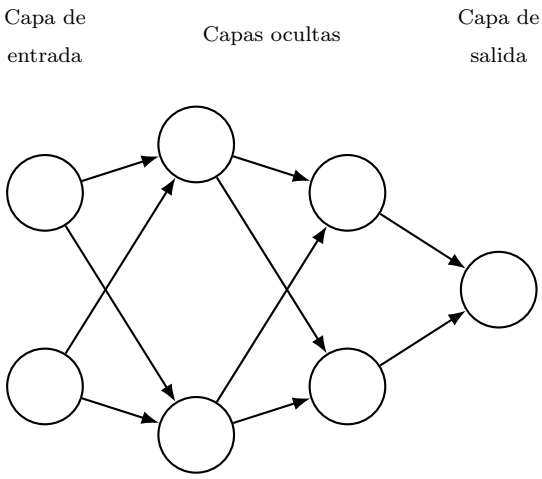


Figura 2.1: Arquitectura básica de una ANN.

2.4.1. Redes neuronales profundas

Cuando una ANN cuenta con múltiples capas, se considera un modelo de DL. Las redes neuronales profundas superan muchas de las limitaciones de los modelos de ML convencionales ya que permiten extraer automáticamente representaciones relevantes directamente de los datos, capturan relaciones complejas e interacciones entre múltiples variables, no requiere transformar las variables a formatos específicos para su procesamiento, y pueden procesar datos mutlidimensionales como los de RMf, entre otros (Sánchez Fernández & Peters 2023, Zhao et al. 2023).

Por ejemplo, Abrol et al. (2021) compararon modelos de DL y ML convencional aplicados a tareas de clasificación y regresión con datos de RM. Sus resultados mostraron que los modelos de DL ofrecieron una mayor precisión y escalabilidad, una extracción de características más eficaz y una mayor capacidad para manejar datos complejos y no lineales, entre otras ventajas.

2.4.2. Modelos especializados: CNNs y RNNs

Existen dos modelos de DL que son de especial utilidad en neuroimagen: las redes neuronales convolucionales (CNN, por sus siglas en inglés) y las redes neuronales recurrentes (RNN, por sus siglas en inglés).

Las CNN están específicamente diseñadas para procesar datos multidimensionales en los que los valores locales están altamente correlacionados, como ocurre con las imágenes (Yan et al. 2022). Las dos principales operaciones que diferencian a las CNN de otros modelos de DL son la convolución y el *pooling*. La operación de convolución permite extraer características de la entrada manteniendo las relaciones espaciales y la operación de *pooling* reduce la dimensionalidad conservando la información más relevante. Tras una concatenación de varias capas de convolución y *pooling*, que extraen desde características simples hasta representaciones más complejas y abstractas, el resultado se aplana y se introduce en una red totalmente conectada que desemboca en la capa de salida. La estructura básica de una CNN puede observarse en la Figura 2.2.

Por ejemplo, Qin et al. (2022) desarrollaron una CNN en grafos utilizando datos de RMf en reposo del consorcio Rest-meta-MDD para clasificar pacientes con y sin diagnóstico de depresión. El modelo alcanzó una precisión del 81,5 % e identificó regiones cerebrales relevantes, como la RMD.

Por el otro lado, las RNNs están específicamente diseñadas para procesar series temporales. Estas redes analizan un elemento a la vez, manteniendo en sus unidades ocultas un vector de estado que contiene información implícita sobre la secuencia completa de datos anteriores (Yan et al. 2022). La estructura básica de una RNN puede observarse en la Figura 2.3.

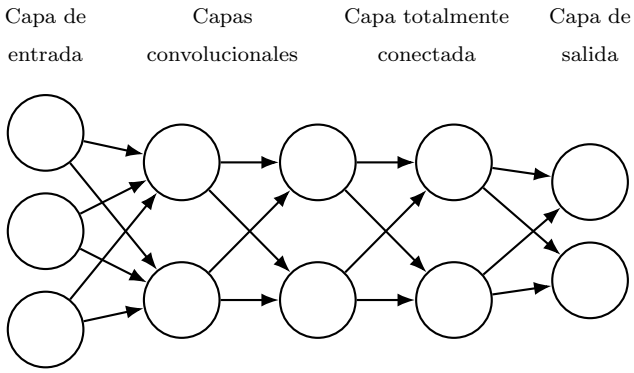


Figura 2.2: Arquitectura básica de una CNN.

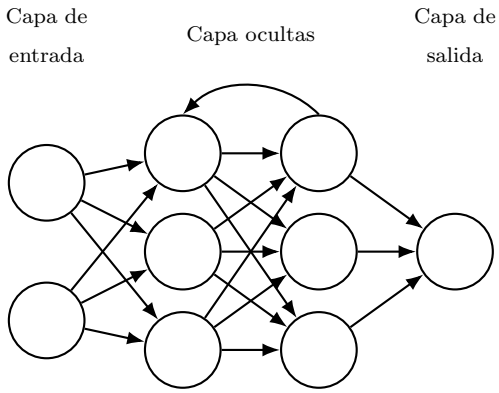


Figura 2.3: Arquitectura básica de una RNN.

Por ejemplo, Sujatha et al. (2021) desarrollaron una RNN utilizando datos de RMf en reposo para clasificar personas con y sin diagnóstico de esquizofrenia, alcanzando una precisión del 81,3 %.

2.4.3. Modelos híbridos CNN-RNN

En estudios que utilizan datos de RMf, ambos modelos resultan especialmente útiles, ya que este tipo de datos se compone de series temporales de imágenes. Es decir, las CNNs procesan la información espacial de las imágenes, mientras que las RNNs modelan la dimensión temporal, permitiendo así el análisis conjunto de la estructura y la función cerebral. Por esta razón, estudios más recientes han optado por una combinación de ambos modelos.

Por ejemplo, Mao et al. (2019) desarrollaron un modelo que combinaba CNNs y RNNs utilizando datos de RMf para clasificar personas con y sin diagnóstico de trastorno por déficit de atención e hiperactividad (TDAH). En su enfoque, las CNNs extrajeron la información espacial de cada imagen de RMf y esta fue posteriormente procesada por una RNN para capturar las dependencias temporales. El modelo logró una precisión del 71,3 %.

En conclusión, la combinación de CNNs y RNNs tiene un gran potencial y ha demostrado superar otras alternativas en la identificación de biomarcadores cerebrales ya que permite analizar conjuntamente la estructura y el funcionamiento del cerebro (Avberšek & Repovš 2022). No obstante, dado que la aplicación de este tipo de modelos en estudios de neuroimagen aún se encuentra en fases iniciales, no existen estudios que hayan aplicado este enfoque combinado al estudio de la depresión.

2.5. Superando las limitaciones del *Deep Learning*

A pesar de las ventajas que los modelos de DL presentan, también tienen una serie de limitaciones. Según Smucny, Shi & Davidson (2022), existen tres limitaciones principales. En primer lugar, la selección de características de datos multidimensionales, como los obtenidos por RMf, representa un gran desafío ya que el entrenamiento de los modelos puede prolongarse durante días e incluso semanas. En segundo lugar, los modelos de DL requieren grandes conjuntos de datos para su entrenamiento, lo cual es difícil de obtener debido al alto coste y tiempo necesario para realizar resonancias magnéticas. En tercer lugar, los modelos de DL han sido comúnmente definidos como "cajas negras", ya que realizan predicciones sin proporcionar información sobre qué características específicas han influido en sus decisiones. Este aspecto es crucial, ya que conocer qué biomarcadores ha identificado el modelo permite desarrollar tratamientos específicos y comprender mejor la naturaleza de los trastornos estudiados.

Por estas razones, recientes estudios han incorporado distintas estrategias para mitigar estas limitaciones. Entre ellas destacan: *transfer learning* para afrontar el problema de la alta dimensionalidad, *data augmentation* para suplir la escasez de datos y *explainable artificial intelligence (XAI)* para abordar la falta de interpretabilidad en los modelos (Smucny, Shi & Davidson 2022).

2.5.1. *Transfer learning*

El *transfer learning* es una técnica que permite aplicar el conocimiento adquirido en la resolución de una tarea a una nueva tarea similar. En DL esto implica reutilizar los parámetros de un modelo previamente entrenado como base para otro modelo que resolverá una tarea similar. Existen distintas variantes de *transfer learning* como *domain adaptation* o *task transfer*, pero todas requieren que los modelos compartan un dominio similar. Es decir, en el campo de la neuroimagen, ambos modelos deberían de entrenarse sobre datos de la misma tipología, como RMf. Otra opción consiste en reutilizar las primeras capas de una CNN, las cuales obtienen características muy genéricas, como bordes, contrastes y texturas, y ajustarlas para la nueva tarea.

Por ejemplo, Dufumier et al. (2024) observaron que los modelos de DL preentrenados sobre grandes bases de datos de población sana superan significativamente a los modelos convencionales en tareas de diagnóstico psiquiátrico.

2.5.2. *Data augmentation*

El *data augmentation* es una técnica que permite generar nuevas instancias a partir del conjunto de datos original mediante la aplicación de diferentes transformaciones, con el objetivo de aumentar su tamaño.

Como Mumuni & Mumuni (2022) indican, una de las estrategias más recientes y con mayor potencial es el *mixup*, la cual consiste en generar nuevas instancias a partir del promedio ponderado de dos muestras aleatorias y sus respectivas etiquetas. En el contexto de la neuroimagen, esto supone combinar datos de RMf de por ejemplo dos pacientes, uno con depresión y otro sano.

Por ejemplo, Smucny, Shi, Lesh, Carter & Davidson (2022) observaron que el rendimiento de una CNN clasificadora entrenada con datos de RMf mejoró significativamente su rendimiento cuando se aplicó *mixup*, aumentando la precisión de un 76,3 % a un 80,4 %.

2.5.3. *Explainable Artificial Intelligence* (XAI)

Las XAI son un conjunto de técnicas que permiten que los modelos de DL no solo proporcionen métricas de precisión, sino también información sobre qué variables han utilizado durante el entrenamiento y la toma de decisiones. En el campo de la neuroimagen, este es un aspecto crítico para la validación clínica de los modelos, ya que resulta imprescindible entender cómo estos han tomado sus decisiones y si dichas decisiones se alinean con el conocimiento previo.

Según diferentes fuentes, estas técnicas pueden clasificarse de diversas maneras. Por ejemplo, Bhati et al. (2024) distinguen las siguientes:

- Técnicas basadas en gradientes: utilización de derivadas del modelo para estimar la importancia de las características.
- Técnicas basadas en perturbaciones: evaluación de la modificación de partes de la entrada.
- Técnicas basadas en descomposición: redistribución de la predicción hacia las características de entrada.
- Modelos interpretables por diseño: incorporación de mecanismos interpretables en la arquitectura.

Por su parte, Van der Velden et al. (2022) se basan en el momento de aplicación, clasificando las técnicas como *model-based* (integradas en el diseño del modelo) o *post-hoc* (aplicadas tras el entrenamiento), y en el grado de dependencia del modelo, distinguiendo entre métodos *model-specific* (dependientes de la arquitectura) o *model-agnostic* (no dependientes de la arquitectura). Además, Farahani et al. (2022) proponen categorizar las técnicas XAI según el objetivo de la explicación, diferenciando entre explicaciones funcionales (que describen el comportamiento global del modelo), arquetípicas (que representan instancias prototípicas de una clase) y de relevancia *post-hoc* (que asignan importancia a las características de entrada en una predicción concreta). Estas múltiples clasificaciones reflejan la diversidad de enfoques en XAI y la necesidad de seleccionar cuidadosamente la técnica más adecuada según el contexto y los objetivos del estudio.

Hasta la fecha, no se han reportado técnicas XAI específicas o ampliamente estandarizadas para el tratamiento de datos de RMf, y las aplicaciones existentes son adaptaciones de métodos desarrollados inicialmente para datos estáticos (imágenes). Aunque Thomas et al. (2019) introducen el modelo DeepLight, que combina CNN+LSTM con LRP aplicado sobre secuencias de RMf, este requiere modificar la arquitectura del modelo.

Integrated Gradients (Sundararajan et al. 2017) destaca como una técnica XAI que permite mantener la arquitectura original sin modificaciones y adaptar el análisis interpretativo a datos RMf secuenciales de forma eficiente y flexible. Más específicamente, IG es un método *post-hoc*, ya que se aplica tras el entrenamiento del modelo; *model-specific*, porque requiere acceso al grafo computacional y a los gradientes; y pertenece al grupo de técnicas basadas en gradientes, al estimar la relevancia de cada característica de entrada mediante la integración de derivadas a lo largo de una trayectoria desde una entrada de referencia (*baseline*) hasta la entrada real.

Por ejemplo, Kim (2025) desarrollaron un método para estimar la conectividad cerebral efectiva (en datos simulados, MEG y RMf) mediante redes LSTM combinadas con IG. Los resultados muestran que el enfoque logra una alta precisión en la detección de conexiones causales (hasta un 96 % en datos simulados) y ofrece mapas interpretables que reflejan relaciones significativas entre regiones cerebrales.

En conclusión, las técnicas XAI aumentan la fiabilidad de los modelos de DL, ya que permiten verificar si las regiones identificadas coinciden con hallazgos previos en la literatura. Además, facilitan la exploración de nuevas regiones cerebrales que podrían haber pasado desapercibidas en estudios anteriores.

2.6. El desafío de los *noisy labels*

Otra importante limitación que los modelos de DL presentan es su sensibilidad a los *noisy labels*. En el contexto de la neuroimagen, esto puede implicar por ejemplo que los datos de una persona diagnosticada con depresión no correspondan realmente a dicha condición y viceversa. Esto conlleva que el modelo se entrene información incorrecta y en consecuencia llegue a conclusiones erróneas. Este problema es especialmente relevante en el campo de los trastornos psiquiátricos como la depresión ya que el porcentaje de diagnósticos erróneos puede alcanzar hasta más de un tercio de los casos en algunos trastornos (Ayano et al. 2021).

Debido al impacto significativo que puede tener este tipo de ruido en los datos, numerosos estudios han propuesto mecanismos para mitigar sus efectos en el entrenamiento de los modelos de DL (Song et al. 2022).

Según Karimi et al. (2020), las distintas técnicas para la reducción y gestión de los *noisy labels* se pueden clasificar en las siguientes seis categorías:

- Limpieza y preprocesamiento de etiquetas: corrección o eliminación de los datos mal etiquetados antes del entrenamiento.
- Arquitectura de la red: incorporación de mecanismos que modelen explícitamente la probabilidad del error en las etiquetas.
- Funciones de pérdida: diseño de funciones de pérdida menos sensibles a los *noisy labels*.
- Reponderación de datos: asignación de pesos a los datos según su probabilidad de ser erróneos.
- Consistencia de datos y etiquetas: cálculo de la similitud entre los datos para detectar *noisy labels*.
- Procedimientos de entrenamiento: uso de procedimientos de entrenamiento alternativos para evitar el sobreajuste a los *noisy labels*.

Por ejemplo Han et al. (2018) aplicaron la técnica de *co-teaching* en el entrenamiento de un modelo de DL. Esta técnica consiste en entrenar dos modelos en paralelo y seleccionar, en cada mini-lote, las instancias con menor pérdida. Estas instancias se pasan a la otra red para que las use en su actualización, lo cual reduce la propagación del error. Esta técnica demostró una mejora significativa en la precisión frente a otros métodos en escenarios con alto nivel de ruido.

También, Lee et al. (2018) utilizarón la técnica *CleanNet* para reducir el número de *noisy labels* en los datos. Esta técnica consiste en una ANN que aprende representaciones para cada

clase a partir de un conjunto de datos de referencia y compara el resto de los datos con estas referencias para detectar *noisy labels*. Mediante esta estrategia, lograron reducir en un 41,5 % el número de *noisy labels* en clases no verificadas, lo que demuestra su efectividad con muy poca supervisión.

En conclusión, existen diversas estrategias para gestionar *noisy labels* en modelos de DL. Sin embargo, su desarrollo es relativamente reciente y su aplicación específica al ámbito de la neuroimagen y el estudio de la depresión aún no ha sido explorada en profundidad.

2.7. Síntesis

A pesar de los importantes avances en la comprensión de la estructura y el funcionamiento cerebral asociados a la depresión mediante el uso de técnicas de ML aplicadas a la neuroimagen, los biomarcadores cerebrales específicos aún no están claramente definidos. Modelos más recientes de DL como las CNN y las RNN han mostrado resultados prometedores en otras condiciones psiquiátricas, pero su aplicación en el estudio de la depresión es todavía muy limitada. Además, estos modelos presentan limitaciones técnicas que comienzan a ser abordadas mediante nuevas estrategias metodológicas.

2.8. Conclusión

El presente proyecto tiene como objetivo contribuir al avance en este campo mediante el desarrollo de un modelo novedoso de DL aplicado al análisis de datos de neuroimagen en depresión. Para ello, se empleará una arquitectura híbrida CNN+RNN que permita para aprovechar simultáneamente la capacidad de las CNN para extraer información espacial y la de las RNN para modelar secuencias temporales.

Con el fin de superar algunas de las limitaciones que presentan estos modelos, se aplicarán las técnicas de *transfer learning* (para reducir el tiempo de entrenamiento del modelo), *data augmentation* (para aumentar la cantidad de datos disponibles), y XAI (para facilitar la identificación de regiones cerebrales relevantes). Además, se aplicarán técnicas para el tratamiento de *noisy labels* para reducir el impacto de las posibles etiquetas diagnosticas potencialmente erróneas.

Se trata por lo tanto, de una metodología novedosa e integral, que hasta donde alcanza el conocimiento actual, no ha sido aplicada previamente al estudio de los biomarcadores de la depresión mediante datos de RMf.

Capítulo 3

Metodología y desarrollo técnico

En este capítulo se describen los recursos utilizados y la metodología implementada para el desarrollo del proyecto. En concreto, se detallan las características del conjunto de datos empleado, las herramientas utilizadas, la estrategia metodológica seguida y las técnicas de preprocesamiento de datos aplicadas. Además, se incluye una sección dedicada a las estrategias y técnicas que no ofrecieron buenos resultados o que no pudieron aplicarse por diversas limitaciones.

3.1. Fuente y descripción del conjunto de datos

El conjunto de datos empleado provino del estudio de Bezmaternykh et al. (2021). Los datos consisten en imágenes de RMf en reposo adquiridas con un escáner Philips Ingenia 3T utilizando una secuencia de imágenes ecoplanares con los siguientes parámetros: tamaño de vóxel = 2×2×5 milímetros, tiempo de repetición = 2500 milisegundos, tiempo de eco = 35 milisegundos, 25 cortes por volumen y un total de 100 volúmenes por participante. Los datos se encuentran en formato NIfTI (.nii.gz) y se distribuyen en acceso abierto junto con archivos de metadatos que contienen información demográfica y clínica de los participantes. En la Figura 3.1 se puede observar un corte axial central de uno de los participantes en el estudio.

La muestra estuvo compuesta por 72 participantes, incluyendo personas con y sin diagnóstico de depresión leve o moderada. Entre las variables clínicas utilizadas se encontraban varias escalas psicométricas empleadas para evaluar la sintomatología depresiva y otros factores psicológicos relevantes, como la Montgomery–Åsberg Depression Rating Scale (Montgomery & Åsberg 1979) o el Beck Depression Inventory (Beck et al. 1961), entre otras. Estas métricas fueron utilizadas en el estudio original para caracterizar la severidad del cuadro clínico, pero no se emplearon directamente en los modelos desarrollados. El rango de edad media de los participantes fue de 33.1 ± 9.5 años en las personas diagnosticadas con depresión y de 33.8 ± 8.5 años

en las personas sin diagnóstico de depresión. Como se muestra en la Figura 3.2, la mayoría de los participantes padecían depresión (51 frente a 21 sin diagnóstico). Además, también existía una desproporción en cuanto al sexo biológico, siendo las personas de sexo biológico femenino más del doble (53 frente a 19 personas de sexo biológico masculino).

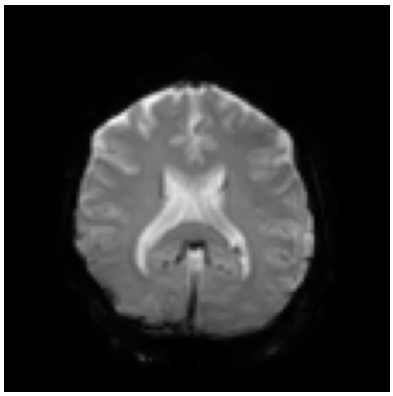


Figura 3.1: Corte axial central de un participante.

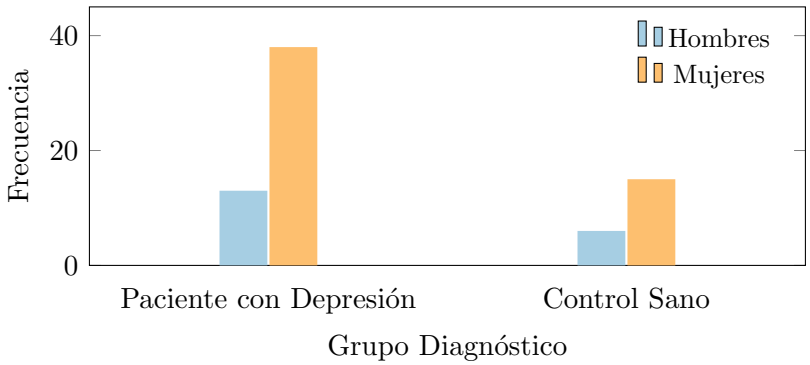


Figura 3.2: Distribución de participantes por grupo diagnóstico y género.

3.2. Entorno computacional y herramientas utilizadas

La implementación del proyecto se realizó utilizando *Python 3* debido a su amplia disponibilidad de librerías especializadas en el procesamiento de neuroimagen y el desarrollo de modelos de ML. En particular, las principales librerías utilizadas se pueden observar en el siguiente listado y en Tabla 3.1:

- *nilearn.masking* y *nibabel*: carga, visualización y enmascaramiento de los datos de RMf en formato NIfTI.
- *tensorflow.keras.layers*: definición de arquitecturas de CNN y RNN.

- *tensorflow.keras.applications*: incorporación de modelos preentrenados mediante técnicas de *transfer learning*.
- *cv2*: tareas de *data augmentation*.
- *tensorflow.keras.losses*: implementación de técnicas de manejo de *noisy labels*.
- *optuna*: optimización de hiperparámetros.
- *codecarbon*: cálculo de emisiones de CO_2 de los modelos.

Estas librerías representan únicamente los principales recursos utilizados durante el desarrollo del proyecto. La lista completa de dependencias, junto con sus versiones específicas, puede consultarse en el archivo *requirements.txt*, disponible en el repositorio del proyecto (Chivite 2025).

Tarea	**Librería**
Visualización y enmascaramiento fMRI	*nilearn.masking*
Normalización y manipulación de imágenes	*nibabel*
Implementación de CNN y RNN	*tensorflow.keras.layers*
Transfer Learning	*tensorflow.keras.applications*
Data Augmentation	*cv2*
Manejo de etiquetas ruidosas	*tensorflow.keras.losses*
Ajuste fino de hiperparámetros	*optuna*
Cálculo de emisiones de CO_2	*codecarbon*

Tabla 3.1: Principales librerías utilizadas durante el desarrollo de las diferentes fases del proyecto.

El desarrollo se llevó a cabo en el entorno de ejecución *Google Colaboratory*, más específicamente su versión *Pro*, mediante *notebooks* diferenciados para cada fase del proyecto. Los archivos de datos y resultados fueron gestionados a través de *Google Drive*.

Para las tareas no relacionadas con el entrenamiento de modelos de DL, se utilizó el entorno de ejecución por CPU con alta disponibilidad de memoria RAM. Por su parte, todos los modelos de redes neuronales fueron entrenados utilizando la GPU NVIDIA A100 proporcionada por *Google Colab*.

El código fuente del proyecto, excluyendo los archivos de particionado del conjunto de datos debido a su tamaño, se encuentra disponible en el repositorio público del proyecto en la plataforma *GitHub* (Chivite 2025).

3.3. Estrategia metodológica y diseño experimental

La metodología del presente proyecto se fundamentó en una serie de estudios clave que abordan los desafíos actuales en el uso de DL para el análisis de datos de RMf. En la Figura 3.3 y la Tabla 3.2 se detalla la estructura concreta de esta metodología y su implementación progresiva.

Como se ha mencionado anteriormente, los datos de RMf contienen información tanto espacial como temporal. Por esta razón, en la etapa 3 (Figura 3.3) se adoptó una arquitectura híbrida CNN (para el análisis espacial) y RNN (para capturar la dinámica temporal). Concretamente, se empleó una *Gated Recurrent Unit* (GRU), una variante simplificada de las RNN que permite reducir los tiempos de entrenamiento sin comprometer el rendimiento, en comparación con alternativas más complejas como las redes LSTM (*Long Short-Term Memory*) (Mienye et al. 2024).

Diversos estudios, como los de Davatzikos (2019) y Janssen et al. (2018), identificaron limitaciones comunes en el uso de técnicas de ML aplicadas a neuroimagen, incluyendo tamaños muestrales reducidos, opacidad en la interpretación de los modelos, presencia de etiquetas erróneas en los conjuntos de datos, y la naturaleza de "caja negra" de muchos modelos de DL.

Con el objetivo de mitigar estas limitaciones, Smucny, Shi & Davidson (2022) propusieron una serie de estrategias: el uso de *data augmentation* y *transfer learning* para contrarrestar la escasez de datos, así como técnicas de XAI para mejorar la interpretabilidad de los modelos.

La estrategia de *data augmentation* que se implementó en la etapa 4 (Figura 3.3) se basó en el enfoque de Krishnapriya & Karuna (2023), la cual consiste en aplicar transformaciones geométricas simples sobre las imágenes de entrada, tales como rotaciones, desplazamientos y duplicación de ejes, con el fin de aumentar artificialmente la variabilidad del conjunto de datos y mejorar la generalización de los modelos.

La estrategia de *transfer learning* que se adoptó en la etapa 5 (Figura 3.3) también sigue la propuesta de Krishnapriya & Karuna (2023), quienes implementaron distintos modelos CNN preentrenados sobre bases de datos de imágenes. Más específicamente, los modelos utilizados en su estudio incluyeron VGG-16, VGG-19, ResNet50 e Inception V3. En el presente proyecto se replicaron los tres primeros, sustituyendo Inception V3 por MobileNetV2, una arquitectura más ligera que permite reducir la carga computacional, especialmente útil en contextos con recursos limitados y conjuntos de datos reducidos. En cuanto al modo de transferencia, se optó por congelar todas las capas convolucionales y re-entrenar únicamente la capa completamente conectada final, siguiendo también el procedimiento descrito en el estudio mencionado.

Con el objetivo de mejorar la robustez de los modelos preentrenados desarrollados, en la etapa 6 (Figura 3.3) se exploraron diferentes variantes estructurales: congelación parcial de capas convolucionales (Ardalan & Subbian 2022), mecanismos de atención (Al Olaimat et al.

2025), y *slice stacking* (Avesta et al. 2023).

Para abordar la posible presencia de *noisy labels* en los datos, en la etapa 7 (Figura 3.3) se evaluaron distintas funciones de pérdida robustas, de acuerdo con las recomendaciones de Karimi et al. (2020). Asimismo, en la etapa 8 (Figura 3.3) se realizó una optimización de hiperparámetros mediante búsqueda automatizada, una etapa fundamental en el diseño y evaluación de modelos de ML (Bischl et al. 2021).

Respecto a la interpretabilidad de los modelos de la etapa 9 (Figura 3.3), se implementó la técnica de *integrated gradients* con el objetivo de estimar la contribución de cada píxel al resultado de la predicción. Esta técnica permite identificar las regiones anatómicas más relevantes dentro de cada corte axial, facilitando así una mejor comprensión de los patrones espaciales y temporales que el modelo considera durante la toma de decisiones.

En resumen, la metodología desarrollada en este proyecto consistió en la aplicación incremental y combinada de las estrategias mencionadas sobre un modelo CNN+GRU entrenado con datos de RMf de personas con y sin diagnóstico de depresión.

3.4. Técnicas de preprocesamiento de datos

Como indican Jaber et al. (2019), no existe un consenso estricto sobre las tareas que deben realizarse en la etapa de preprocesamiento, ya que estas tienen distintos efectos sobre los datos de RMf. A pesar de ello, algunas de las más comunes incluyen la corrección del tiempo de adquisición de cortes (*slice timing correction*), la realineación (*motion correction*) y el suavizado espacial (*smoothing*), entre otras.

El estudio de Bezmaternykh et al. (2021), del cual se obtuvo el conjunto de datos, implementó las siguientes tareas de preprocesamiento con SPM12 en MATLAB:

- Eliminación de los primeros 5 volúmenes temporales.
- Corrección de movimiento.
- Corrección de desfase temporal entre cortes.
- Normalización al espacio MNI (Montreal Neurological Institute), con voxelado de 2×2×2 milímetros. Este es un sistema de referencia anatómico estandarizado, desarrollado a partir del promedio de escaneos cerebrales de múltiples individuos (Chau & McIntosh 2005).
- Suavizado con un kernel gaussiano de 8 milímetros.

Nº	Etapa / Experimento	Objetivo principal	Técnicas / Modelos aplicados
1	Exploración de datos	Comprender las características de entrada y la demografía de la muestra	Análisis de la señal BOLD y análisis descriptivo
2	Preprocesamiento del conjunto de datos	Preparar los datos para su uso en redes neuronales	Recorte de volúmenes, máscara cerebral, z-score, selección de slice, conversión a `float32`
3	Desarrollo de modelos básicos	Evaluar el impacto de la dimensión temporal	CNN 2D vs. CNN 2D+GRU
4	Modelo CNN 2D+GRU con *data augmentation*	Mejorar la capacidad de generalización del modelo	Transformaciones geométricas simples (rotación, desplazamiento, duplicación de ejes)
5	Modelos CNN 2D+GRU preentrenados	Evaluar arquitecturas CNN previamente entrenadas	VGG-19, VGG-16, ResNet50, MobileNetV2
6	*Transfer learning* con modificaciones estructurales	Explorar mejoras estructurales sobre el modelo base	Slice stacking, mecanismo de atención, descongelación de capas convolucionales
7	Modelos con manejo *noisy labels*	Incrementar la robustez ante errores en las etiquetas	*Focal Loss*, *iMAE*, *Label Smoothing*
8	Optimización de hiperparámetros	Afinar el rendimiento del modelo más prometedor	Ajuste automático con *Optuna*
9	Técnicas de interpretabilidad	Estimar la contribución de cada píxel al resultado de la predicción	*Integrated gradients*

Tabla 3.2: Resumen de las etapas metodológicas implementadas en el desarrollo del proyecto.

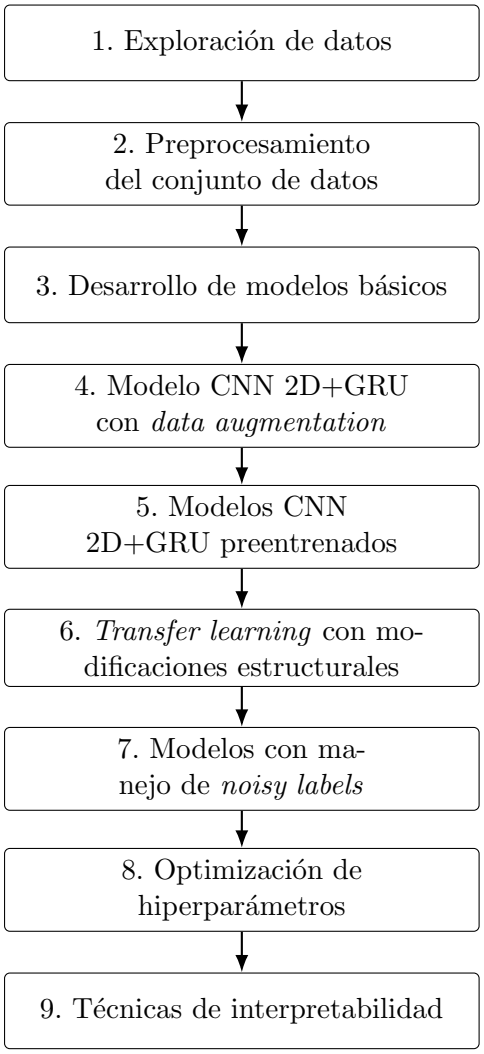


Figura 3.3: Flujo metodológico del proyecto: etapas de desarrollo de modelos y análisis.

En el presente proyecto se optó por desarrollar modelos 2D en lugar de utilizar volúmenes 3D completos por dos razones principales. Primero, debido a las restricciones computacionales presentes en el proyecto. Segundo, debido a que los modelos preentrenados que se implementaron en la fase de *transfer learning* requerían que sus datos de entrada estuvieran en 2D. El desarrollo de modelos 2D es una práctica habitual en estudios de neuroimagen por las mismas razones indicadas anteriormente, menor coste computacional y disponibilidad de modelos preentrenados en 2D Bernal et al. (2019). Cabe destacar que estos modelos 2D (selección de un único corte cerebral) pierden importante información espacial, lo cual es de gran importancia teniendo en cuenta el gran número de interconexiones entre diferentes estructuras cerebrales. A pesar de ello, diferentes estudios sugieren que los modelos en 2D ofrecen igualmente resultados precisos y fiables Avesta et al. (2023). También existen modelos denominados 2.5D los cuales se basan en introducir un mayor número de cortes 2D en diferentes canales del modelo Avesta et al.

(2023).

Por lo tanto, dado que el enfoque que se adoptó en este proyecto se basa en el análisis secuencial de un único corte axial por participante (el corte con mayor varianza), no se aplicó la corrección por tiempo de adquisición de cortes del estudio original ya que este procedimiento está diseñado para alinear temporalmente múltiples cortes dentro de un volumen, lo cual no aplica al tratarse de un único corte. Además, tampoco se aplicó suavizado espacial ya que el suavizado tridimensional implica la introducción de información de cortes adyacentes que no se utilizan en el modelo y el suavizado bidimensional podría diluir detalles espaciales relevantes para el ML. Teniendo esto en cuenta y los resultados obtenidos en la etapa 1 de exploración de datos (Figura 3.3), las tareas de preprocesado que se llevaron a cabo en la etapa 2 (Figura 3.3) del presente proyecto incluyen las siguientes:

- Recorte de volúmenes iniciales: eliminación de los primeros volúmenes de cada sujeto que suelen contener artefactos producidos por la estabilización del escáner o movimientos del participante.
- Aplicación de máscara cerebral: exclusión de voxeles fuera del cerebro (hueso, aire, fondo), conservando únicamente la señal funcional cerebral.
- Normalización *z-score*: estandarización de intensidades por voxel para homogeneizar la señal entre sujetos y regiones.
- Conversión a *float32*: reducción del tamaño en memoria sin pérdida significativa de precisión, optimizando el rendimiento del entrenamiento.
- Selección de un único slice: selección del slice con la mayor varianza.

Una vez completadas las tareas de preprocesamiento descritas, los datos resultantes se almacenaron en formato *.npy*, propio de la librería *NumPy*, con el objetivo de optimizar tanto el tiempo de carga como el rendimiento durante el entrenamiento de los modelos. Esta estrategia permite separar la etapa 2 (Figura 3.3) de preprocesamiento de las etapas 3 y 4 (Figura 3.3) de modelado y facilita la reproducibilidad de los experimentos.

Además del preprocesamiento general, se definieron dos fases adicionales de preparación de datos, adaptadas a los requisitos específicos de los modelos de la etapa 3 (figura 3.3). La metodología que se aplicó sigue las directrices del estudio de Krishnapriya & Karuna (2023), en el que evaluaron diversos modelos CNN preentrenados aplicados a imágenes de RM. En dicho estudio, se aplicaron las siguientes estrategias de *data augmentation*: rotación del 15 %, desplazamiento horizontal y vertical del 5 %, escalado mediante un factor de 1/255 y reflejo según los ejes horizontal y vertical. Con el fin de mantener la coherencia metodológica, en el

presente proyecto se replicaron dichas transformaciones, con la única excepción del re-escalado, que fue omitido. Adicionalmente, con el fin de equilibrar el conjunto de datos original, se aplicó el *data augmentation* con una proporción de ×5 para la clase control y ×2 para la clase depresiva.

Por su parte, en la fase de preparación para *transfer learning*, las imágenes fueron redimensionadas de su tamaño original (112×122 píxeles) a 224×224 píxeles, y se amplió su número de canales a tres mediante replicación (pseudo-RGB), de forma que se ajustaran al formato de entrada requerido por los modelos preentrenados empleados.

3.5. Desarrollo iterativo y toma de decisiones técnicas

A lo largo del desarrollo del presente proyecto, el flujo de trabajo estuvo marcado por una constante revisión y experimentación. Lejos de ser un proceso lineal, el avance se dio por prueba y error, re-visitando continuamente fases anteriores, evaluando alternativas y tomando decisiones basadas en las limitaciones técnicas, computacionales y metodológicas que fueron surgiendo. En esta sección se detalla esta evolución.

3.5.1. Exploración de datos

Originalmente, el proyecto iba a hacer uso del conjunto de datos del consorcio REST-meta-MDD (Yan et al. 2019), un proyecto chino que proporciona datos en abierto de RMf en reposo de 1300 pacientes con trastorno depresivo mayor y 1128 controles sanos de diferentes centros. Sin embargo, a pesar de que la documentación oficial lo describía como datos RMf, el conjunto disponible no incluía volúmenes 4D (espacio + tiempo), sino únicamente mapas derivados (ReHo, ALFF, etc.) ya procesados, sin dimensión temporal. Esto obligó a hacer uso del conjunto de datos de Bezmaternykh et al. (2021), un conjunto de datos significativamente más reducido (72 muestras frente a 2428).

3.5.2. Preprocesamiento del conjunto de datos

Inicialmente se optó por trabajar con datos en formato 1D tras aplicar una máscara cerebral, lo que permitía arquitecturas sencillas y baja carga computacional. Sin embargo, esto limitaba el uso de técnicas posteriores como *transfer learning*, y se obtenían resultados que no permitían una comparación directa con modelos que empleaban un mayor número de dimensiones, por lo que se abandonó.

A pesar del interés inicial por modelos 3D, este enfoque no fue viable ya que no existen modelos 3D preentrenados ampliamente disponibles y el coste computacional de estos modelos

es excesivamente alto.

Por esta razón, se optó por seleccionar un único corte axial del cerebro (el de mayor varianza). Esta decisión supuso una pérdida evidente de información 3D, pero permitió beneficiarse del uso de arquitecturas CNN preentrenadas.

También se probaron diferentes técnicas de normalización: media global frente a *z-score* por sujeto. Finalmente, se optó por la segunda para evitar fuga de información diagnóstica.

La etapa 2 de preprocesamiento (Figura 3.3) se desarrolló de forma modular con el objetivo de poder ser reutilizada para las distintas variaciones posteriores y se re-diseñó su flujo para evitar el almacenamiento redundante de productos intermedios.

3.5.3. Desarrollo de modelos básicos

En esta etapa no se encontraron retos considerables. Se probaron modelos con datos de diferente dimensionalidad para comprobar su viabilidad (1D, 2D y 3D). A pesar de que todos los modelos se sobreajustaban debido al reducido tamaño del conjunto de datos, los modelos 1D y 3D obtuvieron los mejores resultados. Sin embargo, por las razones mencionadas anteriormente, su implementación se excluyó del proyecto.

También se probaron diferentes tamaños de kernel y profundidad. Los modelos más complejos obtuvieron los peores resultados, de nuevo como consecuencia del reducido conjunto de datos. Por esta razón se optó por un modelo más simple como base.

3.5.4. Modelo CNN 2D+GRU con data augmentation

La mayor problemática en esta sección se debió a la correcta selección de técnicas de *data augmentation*. Se intentó la implementación de técnicas más recientes y complejas como el *mixup* mencionado anteriormente (Mumuni & Mumuni 2022). Sin embargo, esta técnica resultó ser demasiado compleja para este proyecto y su implementación requería además de una mayor capacidad computacional. Por esta razón, siguiendo los pasos de Krishnapriya & Karuna (2023), se implementaron una serie de transformaciones geométricas simples que no requerían alta carga computacional.

3.5.5. Modelos CNN 2D+GRU preentrenados

Eventualmente, se encontró un estudio que proporcionaba una implementación de modelos 3D preentrenados para uso en imágenes médicas (Solovyev et al. 2022). Se intentó adaptar dicha implementación con el objetivo de aprovechar la dimensionalidad 3D del conjunto de datos y evitar la pérdida de información espacial. Sin embargo, debido a la complejidad de la adaptación y al elevado coste computacional, se continuó con la implementación 2D.

Originalmente, los modelos a desarrollar fueron VGG-19, VGG-16 y ResNet50, según los resultados del estudio de Krishnapriya & Karuna (2023). Sin embargo, dado que ninguno de ellos obtuvo resultados suficientemente positivos, se optó por implementar también MobileNetV2, siendo este el que ofreció finalmente los mejores resultados.

Otra dificultad encontrada en esta sección fue la adaptación del conjunto de datos preprocesado al formato requerido por los modelos preentrenados. Por esta razón, se reestructuró la etapa 2 (Figura 3.3) de preprocesamiento para que fuera adaptable tanto a esta fase como a la fase de *data augmentation*.

3.5.6. *Transfer learning* con modificaciones estructurales

Esta fase no estaba prevista originalmente, pero surgió al no encontrarse un modelo estable con buenos resultados, hasta que se probó MobileNetV2. Se exploraron las variantes estructurales mencionadas anteriormente: congelación de capas convolucionales, mecanismos de atención, y *slice stacking*. Finalmente, dado que los resultados no superaban a los obtenidos por el modelo MobileNetV2+GRU, se seleccionó este como modelo final por su estabilidad y eficiencia.

3.5.7. Optimización de hiperparámetros

Se identificó que el tamaño de los lotes en la creación de los *tf.data.Dataset* debía mantenerse en 2 para evitar errores de tipo *Out of Memory* (OOM) en Colab.

3.5.8. Modelos con manejo de etiquetas ruidosas

Se implementaron *co-teaching* y validación cruzada, pero la memoria RAM del entorno era insuficiente. Por lo tanto, ninguna de estas dos técnicas se incluyó finalmente.

3.5.9. Técnicas de interpretabilidad

Se implementó una versión de *Grad-CAM* adaptada al modelo y al conjunto de datos, así como una versión de *occlusion* adaptada a la dimensión temporal. Ninguna de estas implementaciones ofreció buenos resultados, por lo que se optó por una aproximación más simple mediante *Integrated Gradients*, que sí ofreció mejores resultados y se comentará más adelante.

3.5.10. Limitaciones del entorno TensorFlow/Colab

La última versión de TensorFlow (2.19) presenta dificultades a la hora de guardar modelos desarrollados mediante *subclassing*. Todos los modelos tuvieron que ser reentrenados siguiendo

las indicaciones del propio François Chollet, el autor de Keras (Chollet 2025). Esto incluía guardar únicamente los pesos y el historial de entrenamiento. Para reutilizarlos, era necesario definir nuevamente la arquitectura del modelo, compilarla, y cargar los pesos guardados.

En definitiva, el desarrollo del proyecto ha estado marcado por un proceso iterativo de exploración, adaptación y toma de decisiones técnicas, en el que se han integrado enfoques diversos y se han enfrentado activamente las limitaciones del entorno y los datos disponibles.

A continuación, se presentan los resultados obtenidos a partir de las distintas configuraciones de modelos evaluadas, así como el análisis de su rendimiento.

Capítulo 4

Resultados

En este capítulo se presentan los resultados obtenidos por los distintos modelos desarrollados a lo largo del proyecto.

4.1. Hiperparámetros de los modelos.

Los modelos desarrollados a lo largo del presente proyecto comparten una serie de hiperparámetros base que fueron seleccionados manualmente siguiendo prácticas comunes en la literatura (Yang & Shami 2020, Yu & Zhu 2020), salvo en los casos donde se indica una modificación (modelos con *transfer learning*, modelos con funciones de pérdida robustas, etc.),

En cuanto a las capas de los modelos, se limitó el número de las capas convolucionales a tres para evitar un coste computacional elevado. Además, se escogió un número de unidades ascendente con el objetivo de permitir a la red extraer características más abstractas progresivamente. En el caso de las capas recurrentes, se optó por una única capa GRU con 64 unidades, siguiendo estructuras comunes en estudios previos (Varshney et al. 2025, Mao et al. 2019) y con el objetivo de nuevo de reducir el coste computacional.

El *batch size* se fijó a dos debido a problemas de memoria del entorno en algunos modelos cuando el tamaño era mayor. Por su parte, el número de épocas se estableció en 30 con el fin de permitir un margen suficiente para la convergencia de los modelos pero sin incurrir en costes excesivos. Además, se hizo uso de *early stopping* monitorizando la sensibilidad y con una paciencia de cinco épocas para evitar el sobreajuste.

También, se utilizó la activación final sigmoide ya que los modelos trataban un problema de clasificación binaria (depresión/control). Como estrategia extra de regularización se incorporaron capas de *dropout* tras las capas convolucionales y densas. La Tabla 4.1 resume estos hiperparámetros utilizados.

Hiperparámetro	**Valor**
Optimizador	*Adam*
Tasa de aprendizaje	1e-4
Regularización L2	1e-4
Capas convolucionales (modelos base)	3
Capas GRU	1
Unidades GRU	64
Unidades en capa densa	64
Activación final	*Sigmoid*
Dropout convolucional	0.2
Dropout final	0.5
Número de épocas	30
Batch size	2
Early stopping	*Monitor: recall, patience: 5, restore best weights: True*
Función de pérdida	*Binary crossentropy*
Métricas	Exactitud, precisión, sensibilidad, AUC

Tabla 4.1: Hiperparámetros base comunes a los modelos implementados.

4.2. Resultados de modelos base

El objetivo de esta sección era doble. Por un lado, establecer una base sobre la cual poder comparar el impacto de las técnicas posteriores de aumento de robustez de los modelos. Por otro lado, evaluar si la incorporación de la dimensión temporal aportaría beneficios.

La Tabla 4.2 recoge las métricas obtenidas por ambos modelos base (CNN y CNN+GRU). Ambos modelos alcanzaron una sensibilidad del 100 %, pero una especificidad nula. Es decir, los modelos clasificaron todos los casos como pertenecientes a la clase positiva (depresión). Esta situación puede deberse al desequilibrio presente en los datos (51 pacientes con depresión frente a 21 controles sanos). Sin embargo, dado que la muestra ya era reducida, se optó por no implementar un balance de clases.

A pesar de este sesgo, el modelo CNN+GRU obtuvo un AUC un 21 % superior a la del modelo CNN (0,4167 frente a 0,3432). Esto sugiere que el modelo CNN+GRU presenta una mayor capacidad para separar entre clases, aunque dicha mejora no se traduce aún en una especificidad aceptable debido al desbalance de clases.

Desde el punto de vista computacional, también se observaron mejoras en el modelo CNN+GRU.

A pesar de ser más complejo, requirió un 40 % menos de tiempo de entrenamiento (226,03 segundos frente a 377,01 segundos) y generó casi la mitad de emisiones de CO_2 (0,0053 kg frente a 0,0101 kg).

Métrica	CNN	CNN+GRU
Exactitud	0,7273	0,7273
Precisión	0,7273	0,7273
Sensibilidad	1	1
Especificidad	0,0000	0,0000
F1-score	0,8421	0,8421
AUC	0,3432	**0,4167**
Tiempo (s)	377,01	**226,03**
CO_2 (kg)	0,0101	**0,0053**

Tabla 4.2: Resultados de los modelos base: CNN y CNN+GRU sin técnicas adicionales.

En cuanto a las curvas de entrenamiento en validación de los modelos (Figura 4.1), el modelo CNN mostró una curva de pérdida muy irregular a lo largo de las 30 épocas. Esta situación indica un entrenamiento inestable, lo que demuestra una falta de consistencia en las predicciones. En cambio, el modelo CNN+GRU mostró una curva más estable tanto en pérdida como en sensibilidad, y se entrenó durante menos épocas, lo que sugiere una mayor capacidad de generalización y un aprendizaje más robusto.

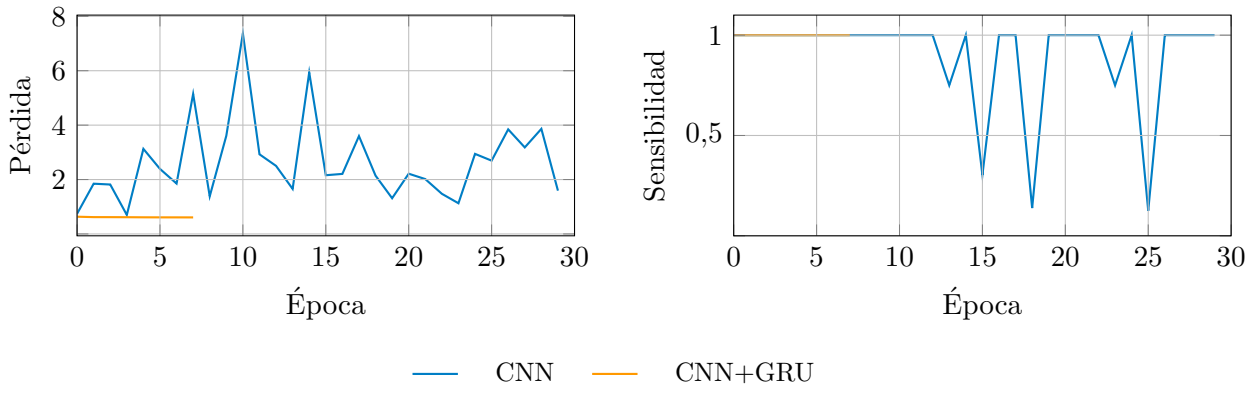


Figura 4.1: Curvas de pérdida y sensibilidad en validación de los modelos base.

En conclusión, aunque ambos modelos fallan en lograr una buena generalización, los resultados del modelo CNN+GRU fueron mejores que los del modelo CNN. Este aspecto sugiere que el modelo se ha beneficiado de la adición de la dimensión temporal.

4.3. Resultados de *data augmentation*

El objetivo de esta sección fue comprobar si el aumento del tamaño de la muestra y su balance mediante la técnica de *data augmentation* resultaba en un modelo con una mayor capacidad de generalización.

La Tabla 4.3 presenta los resultados del modelo CNN+GRU base y del modelo CNN+GRU entrenado con una muestra aumentada y balanceada artificialmente. De nuevo, a pesar del balance de clases, el modelo alcanzó una sensibilidad del 100 %, al igual que el modelo CNN+GRU base. Sin embargo, en el resto de métricas se observó un descenso apreciable, especialmente en precisión (0,5000) y F1-score (0,6667).

Además, desde el punto de vista computacional, el modelo tardó más en entrenarse (377,01 segundos frente a 226,03 segundos) y, por ende, también aumentó la cantidad de CO_2 emitido (0,0101 kg frente a 0,0053 kg). Es decir, no solo se obtuvieron peores resultados, sino que además se incrementó el impacto ambiental.

Métrica	**CNN+GRU (sin *data augmentation*)**	**CNN+GRU (con *data augmentation*)**
Exactitud	**0,7273**	0,5000
Precisión	**0,7273**	0,5000
Sensibilidad	1	1
Especificidad	0	0
F1-score	**0,8421**	0,6667
AUC	**0,4167**	0,4062
Tiempo (s)	**226,03**	377,01
CO_2 (kg)	**0,0053**	0,0101

Tabla 4.3: Comparativa de resultados entre el modelo CNN+GRU base y su versión entrenada con *data augmentation*.

En cuanto a la curva de pérdida en validación (Figura 4.2), el modelo con *data augmentation* mostró una pérdida más alta y plana que el modelo base. Esta situación indica un aprendizaje más lento y menos efectivo.

En conclusión, los resultados sugieren que, en este contexto en particular, la técnica de *data augmentation* no aportó beneficios al rendimiento del modelo. Es posible que esta situación se deba a que el modelo base ya estaba sobreajustado y la nueva muestra no fuera lo suficientemente grande como para eliminar este sobreajuste.

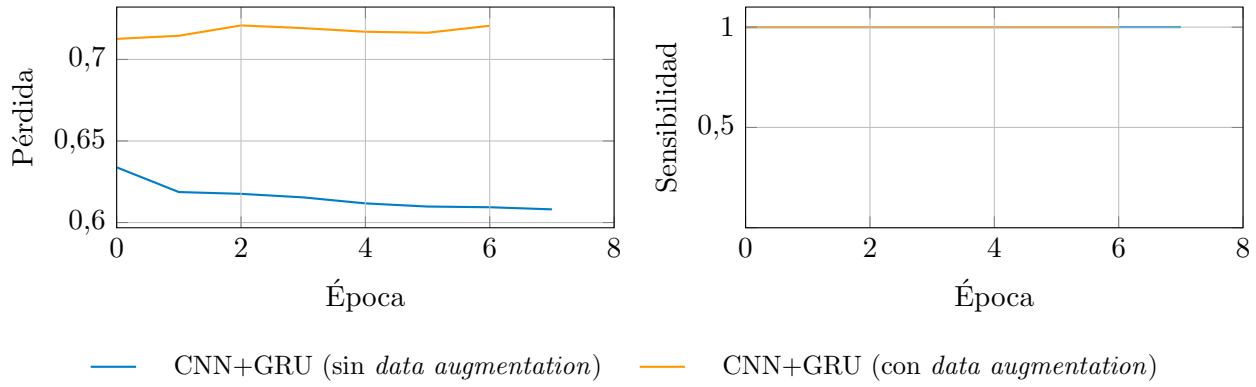


Figura 4.2: Curvas de pérdida y sensibilidad en validación para el modelo CNN+GRU con y sin *data augmentation*.

4.4. Resultados de *transfer learning*

El objetivo de esta sección fue evaluar si la sustitución de las capas convolucionales propias por modelos CNN preentrenados resultaba en una mejora de la capacidad de generalización de estos.

La Tabla 4.4 resume el rendimiento de los cuatro modelos CNN preentrenados (VGG-19, VGG-16, ResNet50 y MobileNetV2) + GRU, comparados con el modelo CNN+GRU con *data augmentation*. Entre todos ellos, MobileNetV2 fue el que alcanzó el mejor rendimiento en todas las métricas. En comparación con el modelo base con *data augmentation*, este modelo obtuvo:

- Una mejora del 71,1 % en la precisión.
- Un aumento del F1-score del 25,8 %.
- Una mejora en el AUC de más del doble (122,1 %).
- Un incremento del 68,8 % en la exactitud.

Además, aunque alcanzó una sensibilidad ligeramente inferior (0,81 frente a 1), este aspecto es positivo, ya que sugiere que el modelo es significativamente más equilibrado y capaz de generalizar. Este equilibrio también se refleja en la especificidad, que pasó de 0 a 0,87.

Los modelos preentrenados con VGG-19 y VGG-16 también mostraron una mejora significativa frente al modelo base. Algunas de estas mejoras fueron:

- La exactitud aumentó un 12,5 % y un 25 %.
- La precisión aumentó un 33,4 % y un 22,2 %.
- La sensibilidad se redujo en un 75 % y un 31,25 %.

- La especificidad aumentó en un 87,5 % y un 56,25 %.

A pesar de no alcanzar los valores de MobileNetV2, VGG-16 fue el modelo más eficiente computacionalmente, requiriendo únicamente 211,64 segundos de entrenamiento (-43,9 % de tiempo) y generando 0,0073 kg de CO_2 (-27,7 % de emisiones). En cambio, MobileNetV2 tuvo un coste computacional mucho más elevado, requiriendo un total de 1462,86 segundos de entrenamiento (+388 %) y generando unas emisiones tres veces mayores que el modelo base (0,0302 frente a 0,0101 kg).

El modelo ResNet50 fue el que obtuvo el peor rendimiento en todas las métricas (salvo en especificidad), además de presentar el mayor coste computacional y de emisiones.

Métrica	**VGG-19 +GRU**	**VGG-16 +GRU**	**ResNet50 +GRU**	**MobileNetV2 +GRU**	**CNN+GRU (con *data augmentation*)**
Exactitud	0,5625	0,6250	0,4375	**0,8438**	0,5000
Precisión	0,6667	0,6111	0	**0,8667**	0.5000
Sensibilidad	0,2500	0,6875	0	0,8125	**1**
Especificidad	**0,8750**	0,5625	**0,8750**	**0,8750**	0
F1-score	0,3636	0,6471	0	**0,8387**	0,6667
AUC	0,6719	0,7266	0,3945	**0,9023**	0,4062
Tiempo (s)	324,75	**211,64**	1287,01	1462,86	377,01
CO_2 (kg)	0,0119	**0,0073**	0,0252	0,0302	0,0101

Tabla 4.4: Comparativa de resultados entre modelos CNN+GRU preentrenados y el modelo CNN+GRU con *data augmentation*

Las curvas de pérdida y sensibilidad en validación (Figura 4.3) también corroboran un mejor entrenamiento de los modelos preentrenados, ya que muestran trayectorias más estables (fíjese en el rango de valores del eje y frente a los modelos anteriores).

En conclusión, los resultados indican que la estrategia de *transfer learning* ofrece resultados favorables. Tres de los cuatro modelos desarrollados han mejorado en todas sus métricas clave y finalmente son capaces de generalizar. MobileNetV2 se consolida como la arquitectura más eficaz, a pesar de su elevado coste computacional, mientras que VGG-16 destaca como una opción intermedia que combina eficiencia energética y buen desempeño.

4.5. Resultados de modificaciones estructurales

Dadas las restricciones computacionales del proyecto, se optó por introducir diversas modificaciones estructurales al modelo preentrenado VGG-16 + GRU, ya que había demostrado un

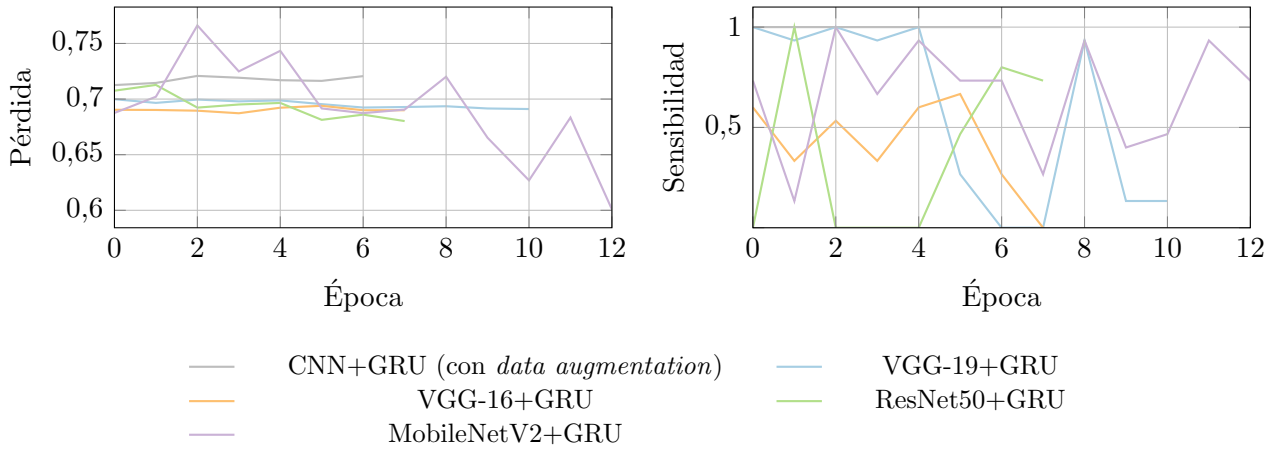


Figura 4.3: Curvas de pérdida y sensibilidad en validación de modelos con y sin *transfer learning*.

buen equilibrio entre rendimiento y eficiencia. Las variantes evaluadas incluyen: *slice stacking*, incorporación de un mecanismo de atención, y descongelado parcial de capas convolucionales.

Tal como se recoge en la Tabla 4.5, cada una de estas modificaciones produjo tanto mejoras como empeoramientos respecto al modelo base:

- El modelo con *slice stacking* redujo el tiempo de entrenamiento un -12,1 % y las emisiones de CO_2 un -17,8 %, convirtiéndose en la variante más eficiente energéticamente. Sin embargo, obtuvo una sensibilidad del 100 % y una especificidad del 0,0625 %, lo cual indica que el modelo es incapaz de generalizar y clasifica todo como perteneciente a la clase positiva (depresión).

- El modelo con mecanismo de atención mejoró la precisión en un +19 % (0,7273 frente a 0,6111) y la especificidad en un +44.4 % (0,8125 frente a 0,5625).

- La variante con capas descongeladas obtuvo el mejor F1-score (+13,1 %) y una sensibilidad un 36,4 % superior a la del modelo base (0,9375 frente a 0,6875), lo que indica una mayor capacidad de detección de positivos manteniendo cierto equilibrio.

Sin embargo, ninguna de las tres modificaciones logró superar al modelo base de forma global. En todas ellas se observaron empeoramientos en el resto de las métricas. Además, solo la variante con *slice stacking* mejoró el coste computacional, las demás variantes aumentaron tanto el tiempo de entrenamiento como las emisiones.

Las curvas de pérdida y sensibilidad en validación (Figura 4.4) muestran que el modelo con capas descongeladas presentó una pérdida más estable y descendente a lo largo de las épocas, indicando un entrenamiento más consistente. En cambio, los modelos con *slice stacking* y mecanismo de atención mostraron mayor oscilación, especialmente en la sensibilidad, lo que sugiere un comportamiento menos robusto y más dependiente de la época de entrenamiento.

Métrica	Slice stacking	Mecanismo de atención	Descongelado de capas	VGG-16 + GRU (base)
Exactitud	0,5312	**0,6562**	**0,6562**	0,6250
Precisión	0,5161	**0,7273**	0,6000	0,6111
Sensibilidad	**1**	0,5000	0,9375	0,6875
Especificidad	0,0625	**0,8125**	0,3750	0,5625
F1-score	0,6809	0,5926	**0,7317**	0,6471
AUC	0,4531	0,6562	0,6328	**0,7266**
Tiempo (s)	**186,02**	341,95	346,19	211,64
CO_2 (kg)	**0,0060**	0,0129	0,0127	0,0073

Tabla 4.5: Comparativa de resultados entre el modelo VGG-16 + GRU base y variantes con modificaciones estructurales.

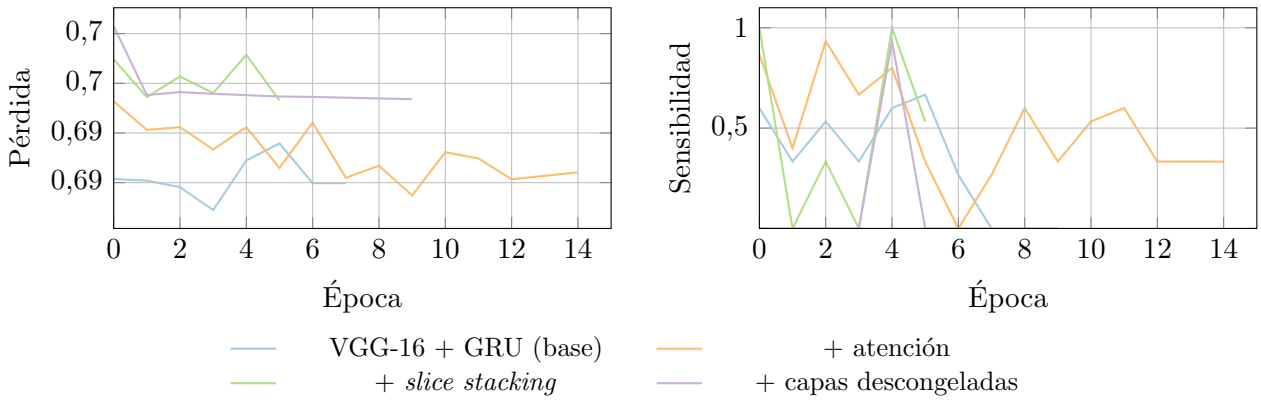


Figura 4.4: Curvas de pérdida y sensibilidad en validación del modelo VGG-16 + GRU con modificaciones estructurales.

En conjunto, los resultados indican que estas modificaciones no ofrecieron mejoras sustanciales frente al modelo base VGG-16+GRU. Por tanto, y dado que ninguna de estas variantes igualó el rendimiento global de MobileNetV2, se mantuvo esta última como la arquitectura de referencia en las siguientes fases del estudio.

4.6. Resultados de funciones de pérdida robustas

El objetivo de esta sección fue examinar si la implementación de diferentes funciones de pérdida robustas podría mejorar el rendimiento del modelo MobileNetV2+GRU, que en las secciones anteriores había alcanzado los mejores resultados en la mayoría de las métricas. Las tres nuevas funciones de pérdida implementadas fueron *Focal Loss*, *Label Smoothing* e *iMAE*.

La Tabla 4.6 recoge los resultados obtenidos por estos modelos:

- El modelo con *Label Smoothing* mantuvo la misma exactitud (0,8438) que el modelo base, pero empeoró en sensibilidad (-7,7 %), F1-score (-1,3 %) y AUC (-2,6 %).

- El modelo con *Focal Loss* ofreció una precisión y especificidad perfectas (+15,4 % y +14,3 % respectivamente frente al modelo base). Sin embargo, la sensibilidad, la exactitud y el AUC se redujeron en un 30,8 %, 7,4 % y 1,3 %, respectivamente.

- El modelo con iMAE empeoró su rendimiento en todas las métricas, reduciendo por ejemplo el *F1-score* y el AUC en un 35 % y un 30,3 %, respectivamente. No obstante, fue el único modelo que mejoró en eficiencia computacional, reduciendo el tiempo de entrenamiento en un 12,4 % y las emisiones de CO_2 en un 9 %.

Métrica	**MobileNetV2 + GRU (base)**	**+ Focal Loss**	**+ Label Smoothing**	**+ iMAE**
Exactitud	**0,8438**	0,7812	0,8438	0,5312
Precisión	0,8667	**1**	0,9231	0,5294
Sensibilidad	**0,8125**	0,5625	0,7500	0,5625
Especificidad	0,8750	**1**	0,9375	0,5000
F1-score	**0,8387**	0,7200	0,8276	0,5455
AUC	0,9023	**0,9141**	0,8789	0,6289
Tiempo (s)	1462,86	1510,75	1445,36	**1281,90**
CO_2 (kg)	0,0302	0,0321	0,0306	**0,0273**

Tabla 4.6: Comparativa entre el modelo MobileNetV2+GRU base y sus variantes con funciones de pérdida robustas.

Las curvas de pérdida y sensibilidad en validación (Figura 4.5) reflejan bien estas diferencias. El modelo con *Focal Loss* presenta una curva de pérdida muy baja y estable, pero su sensibilidad es irregular a lo largo de las épocas. El modelo con *Label Smoothing* muestra un comportamiento más equilibrado y estable, mientras que el modelo con *iMAE*, a pesar de tener la curva de pérdida más baja, no logra mejorar sustancialmente la sensibilidad ni el *F1-score*, lo que indica que esta función de pérdida penaliza en exceso los errores en presencia de etiquetas potencialmente ruidosas.

En conclusión, los resultados sugieren que ninguna de las funciones de pérdida evaluadas mejoró globalmente al modelo MobileNetV2+GRU base. Además, la sensibilidad, que era la métrica prioritaria en este proyecto, disminuyó en los tres casos.

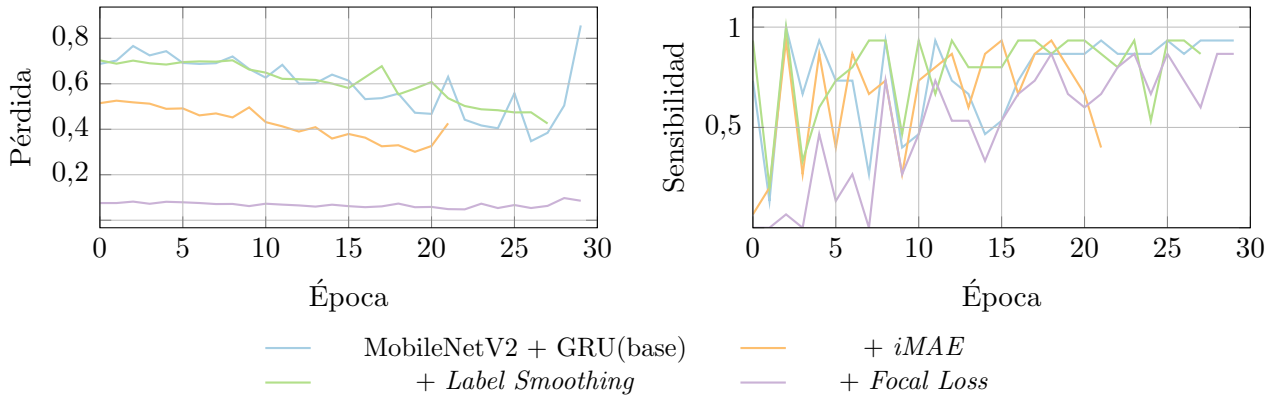


Figura 4.5: Curvas de pérdida y sensibilidad en validación de MobileNetV2 con diferentes funciones de pérdida robustas.

4.7. Resultados de optimización de hiperparámetros.

Con el objetivo de optimizar el rendimiento del modelo con los mejores resultados obtenidos hasta el momento (MobileNetV2+GRU base), se llevó a cabo una búsqueda de hiperparámetros mediante la librería Optuna. En la Tabla 4.7 se indican los hiperparámetros explorados, los cuales incluyen: el coeficiente de regularización L2, el número de unidades en la capa GRU y en la capa densa, la tasa de dropout y el valor de la tasa de aprendizaje. El mejor conjunto de valores obtenidos fue: L2 = 4,94e-05, unidades en la capa GRU = 64, unidades en la capa densa = 64, tasa de *dropout* = 0,289 y tasa de aprendizaje = 0,00099.

Hiperparámetro	**Rango**	**Tipo**
Coeficiente de regularización L2	$[10^{-6}, 10^{-3}]$	Distribución log-uniforme
Unidades GRU	{32, 64, 128}	Valor discreto (categórico)
Unidades en capa densa	{32, 64, 128}	Valor discreto (categórico)
Tasa de *dropout*	[0,2, 0,6]	Valor continuo uniforme
Tasa de aprendizaje	$[10^{-5}, 10^{-3}]$	Distribución log-uniforme

Tabla 4.7: Rango de búsqueda de hiperparámetros utilizado durante la optimización con *Optuna*.

El modelo entrenado con esta configuración logró una sensibilidad del 81,25 %, idéntica a la del modelo MobileNetV2 base. Sin embargo, presentó ligeros descensos en el resto de métricas

(Figura 4.8):

- Exactitud: -3,75 % (0,8125 vs. 0,8438).
- Precisión: -6,3 % (0,8125 vs. 0,8667).
- Especificidad: -7,1 % (0,8125 vs. 0,8750).
- F1-score: -3,1 % (0,8125 vs. 0,8387).
- AUC: -1,3 % (0,8906 vs. 0,9023).

A pesar de esta reducción en métricas globales, el modelo optimizado mejoró notablemente en eficiencia computacional, con una reducción del -14,1 % en el tiempo de entrenamiento (1256,22 s vs. 1462,86 s) y del -52,2 % en las emisiones de CO_2 (0,0144 kg vs. 0,0302 kg).

Métrica	**MobileNetV2 +GRU (base)**	**MobileNetV2 +GRU (optimizado)**
Exactitud	**0,8438**	0,8125
Precisión	**0,8667**	0,8125
Sensibilidad	0,8125	0,8125
Especificidad	**0,8750**	0,8125
F1-score	**0,8387**	0,8125
AUC	**0,9023**	0,8906
Tiempo (s)	1462,86	**1256,22**
CO_2 (kg)	0,0302	**0,0144**

Tabla 4.8: Comparativa entre el modelo MobileNetV2+GRU base y su versión optimizada mediante búsqueda de hiperparámetros.

Las curvas de pérdida y sensibilidad en validación (Figura 4.6) muestran un entrenamiento mucho más estable y progresivo en el modelo optimizado. Su curva de pérdida desciende de forma más pronunciada y constante, alcanzando valores notablemente más bajos que el modelo base. En cuanto a la sensibilidad, el modelo optimizado alcanza antes valores elevados y los mantiene con menor variabilidad a lo largo de las épocas.

En conclusión, a pesar de que el modelo optimizado no logró mejorar las métricas globales, sí permitió reducir el coste computacional y estabilizar el entrenamiento. Sin embargo, dado que el objetivo principal del presente proyecto era obtener un modelo con alta sensibilidad se optó por examinar la interpretabilidad del modelo MobilenetV2 + GRU base en la siguiente sección.

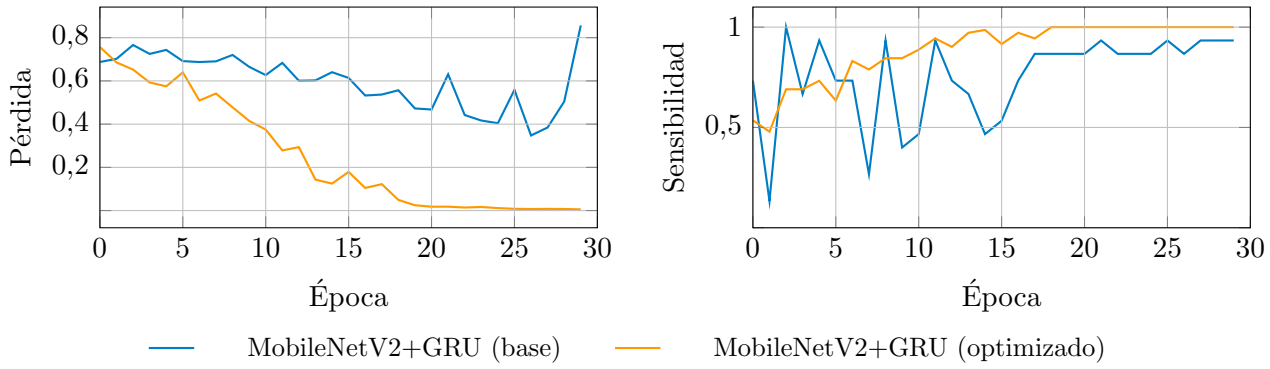
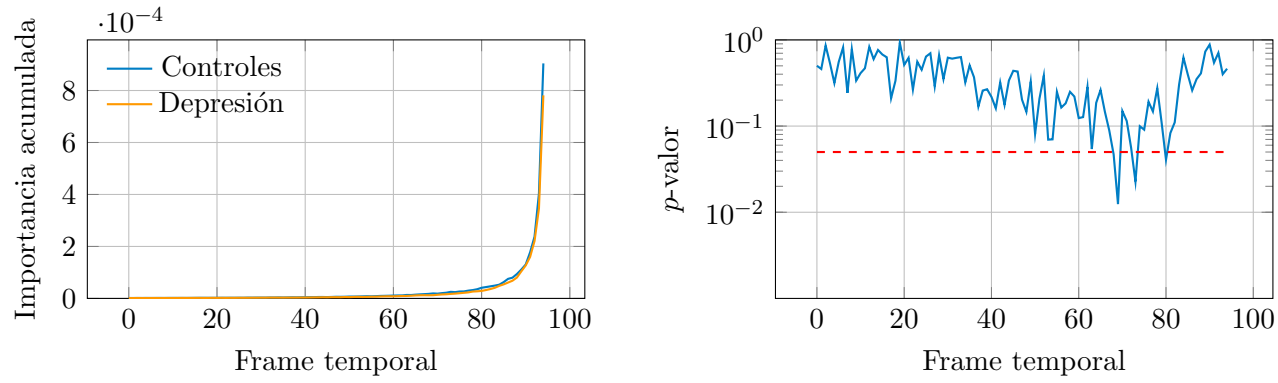


Figura 4.6: Curvas de pérdida y sensibilidad en validación del modelo MobileNetV2 antes y después de la optimización de hiperparámetros.

4.8. Resultados de *Integrated Gradients*

El objetivo de esta sección fue explorar la interpretabilidad del modelo MobilenetV2+GRU mediante *integrated gradients*.

En primer lugar, para examinar la dimensión temporal del modelo, se calculó la importancia temporal media por *frame* mediante el promedio de las atribuciones absolutas por cada instante temporal (Figura 4.7a). Los resultados mostraron que, si bien los últimos *frames* recibían más atribución en promedio, no se observaba una diferencia clara entre clases. Para verificar si existían diferencias significativas no visibles en la media, se aplicó un *test t de Student* por *frame*. El resultado identificó varios *frames* (68, 69, 73 y 80) con $p < 0{,}05$, sugiriendo que esas ventanas temporales podrían ser relevantes en la diferenciación entre clases.



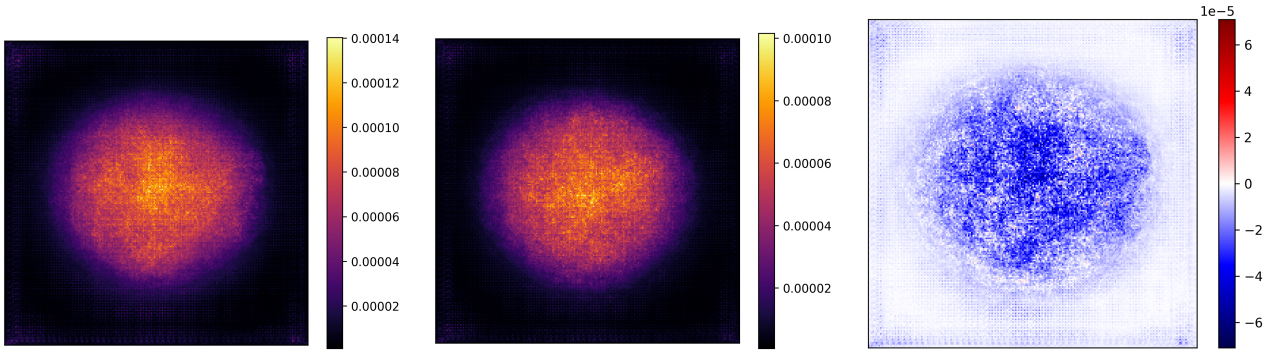
(a) Importancia temporal media por clase.

(b) p-valores por frame (escala logarítmica).

Figura 4.7: Análisis temporal de la importancia de cada *frame*. A la izquierda se muestra la importancia media por clase, y a la derecha los p-valores asociados a la diferencia entre clases.

A partir de estos *frames* significativos, se examinó la dimensión espacial del modelo mediante

el cálculo de los mapas de atribución medios para cada clase (Figuras 4.8a y 4.8b), promediando las atribuciones en los frames seleccionados. También, se calculó la diferencia espacial entre clases (Figura 4.8c). Los resultados mostraron una mayor atribución en sujetos control, especialmente en regiones posteriores y centrales del cerebro. Esto sugiere que el modelo aprende patrones característicos de los controles que están menos presentes en los sujetos con depresión.



(a) Importancia espacial acumulada – clase Control.

(b) Importancia espacial acumulada – clase Depresión.

(c) Diferencia entre clases (control - depresión).

Figura 4.8: Mapas espaciales de importancia media obtenidos mediante *integrated gradients*. Se muestran las regiones más relevantes para cada clase y la diferencia entre ambas.

En conclusión, este análisis de interpretabilidad permitió identificar tanto regiones temporales como espaciales potencialmente discriminativas en la clasificación entre sujetos sanos y con depresión. Aunque las diferencias no fueron marcadamente visibles en el promedio, el análisis estadístico reveló momentos específicos relevantes para el modelo.

4.9. Resultados finales

A lo largo de este capítulo se han presentado los resultados obtenidos para cada uno de los modelos y estrategias implementadas. Se han evaluado modelos base (CNN y CNN+GRU), variantes con *data augmentation*, arquitecturas preentrenadas mediante *transfer learning*, modelos con modificaciones estructurales, variantes con funciones de pérdida robustas y configuraciones optimizadas mediante búsqueda de hiperparámetros. Asimismo, se han incluido métricas relevantes como sensibilidad, exactitud, precisión, F1-score y AUC, así como estimaciones de emisiones de CO_2 y tiempo de entrenamiento.

El modelo que obtuvo el mejor rendimiento general fue el basado en MobileNetV2, destacando por su equilibrio entre sensibilidad (81,25 %), precisión (86,67 %) y AUC (0,9023). El apartado de interpretabilidad del modelo también ofreció resultados interesantes, identificando regiones espaciales y temporales relevantes para la clasificación.

De entre las técnicas implementadas, el *transfer learning* fue la que arrojó los resultados más positivos. En cambio, el *data augmentation*, la implementación de funciones de pérdida robustas y la optimización de hiperparámetros no lograron mejorar globalmente al modelo base. La causa de estos resultados y sus implicaciones se expondrán en el siguiente capítulo.

Capítulo 5

Conclusiones

5.1. Objetivos del estudio y principales hallazgos

Numerosos estudios utilizan modelos de DL sobre datos de neuroimagen para examinar distintas condiciones como el Alzheimer, el Parkinson, la esquizofrenia o el autismo (Chouliaras & O'Brien 2023). Sin embargo, y a pesar de su alta incidencia en la sociedad, su aplicación al estudio de trastornos psicológicos como la depresión sigue siendo escasa. Esta limitación se debe, por un lado, a la falta de comprensión sobre la etiología y fisiopatología de este tipo de trastornos (Prvulovic & Hampel 2010), y por otro, a las barreras técnicas que conllevan estos estudios (requerimiento de grandes conjuntos de datos, escasa interpretabilidad de los modelos, entre otros). Estos factores, generan escepticismo en la comunidad clínica respecto a la utilidad de estos enfoques para el diagnóstico, tratamiento y prevención de trastornos psicológicos. De hecho, la propia Asociación Americana de Psicología desincentiva actualmente su uso en el ámbito clínico (First et al. 2018). No obstante, este estudio ha logrado desarrollar un modelo capaz de clasificar imágenes de RMf de pacientes con depresión y controles sanos con resultados prometedores.

El principal objetivo de este proyecto ha sido desarrollar un modelo con una alta sensibilidad, dado que el modelo tendría aplicaciones en el campo de la salud. Este aspecto se ha conseguido ya que el mejor modelo desarrollado (MovilenetV2+GRU) alcanzó una sensibilidad del 81,25 %, clasificando correctamente 13 de los 16 pacientes con depresión y 14 de los 16 controles sanos. Este hecho sugiere que las personas que padecen depresión presentan diferencias funcionales y/o estructurales en el cerebro, como tantos estudios han sugerido anteriormente (Gray et al. 2020, Castanheira et al. 2019, Schmaal et al. 2020).

Uno de los objetivos secundarios de este proyecto ha sido observar si un modelo híbrido CNN+GRU obtiene mejores resultados que un modelo CNN, dado que según la literatura los pacientes con depresión presentan diferencias tanto estructurales como funcionales en el cerebro

(Castanheira et al. 2019). Sin embargo, los resultados no fueron exactamente los esperados. A pesar de que el modelo CNN+GRU obtuvo un mejor AUC y un mejor uso de los recursos computacionales, en el resto de métricas igualaba al modelo CNN. Esta situación puede deberse al reducido tamaño de la muestra, la cual no permite el desarrollo de modelos complejos.

Otro de los objetivos secundarios del proyecto fue implementar de forma incremental las estrategias propuestas por Smucny, Shi & Davidson (2022) para aumentar la robustez de los modelos (*transfer learning*, *data augmentation* y técnicas XAI). Además, de aplicar diferentes funciones de pérdida robustas para el manejo de *noisy labels* en el conjunto de datos, debido al alto número de diagnósticos erróneos en la práctica clínica (Ayano et al. 2021).

El modelo desarrollado con *data augmentation* obtuvo peores resultados que el mismo modelo sin aumentación. Estos resultados, aunque contradictorios, podrían deberse a que, a pesar de haber triplicado el número de muestras (de 72 a 207), esta cantidad sigue siendo demasiado reducida para desarrollar un modelo de DL desde cero.

Los modelos que utilizaron CNNs preentrenadas sí obtuvieron los resultados esperados y mejoraron significativamente su rendimiento. Pasaron de clasificar los datos de forma errática, incapaces de generalizar, a obtener métricas más estables. Esto demuestra que el *transfer learning* es una técnica imprescindible cuando se cuenta con un conjunto de datos reducido que no permite el desarrollo de modelos desde cero.

De forma contraria a lo esperado, los modelos desarrollados implementando diferentes funciones de pérdida con el objetivo de reducir el impacto de posibles *noisy labels* tampoco obtuvieron mejoras. Esto puede deberse a que, en el estudio del que se obtuvieron los datos, se emplearon hasta 16 escalas psicométricas distintas para diagnosticar a los pacientes, lo que puede haber dado lugar a un etiquetado clínico muy preciso. No obstante, esta no es la situación habitual en la práctica clínica, donde, dependiendo del país, una sola consulta puede ser suficiente para emitir un diagnóstico (Ayano et al. 2021). Por esta razón, a pesar de que no se ha demostrado su utilidad en el presente proyecto, se destaca la importancia de desarrollar modelos robustos frente a *noisy labels* si se pretende aplicar estos modelos en entornos reales, donde los errores diagnósticos son más probables.

Los resultados obtenidos mediante la técnica de IG mostraron que, si bien los últimos *frames temporales* recibían más atribución en promedio, no se observaba una diferencia clara entre sujetos con depresión y controles sanos. Sin embargo, un *Student's t test* aplicado a cada paso temporal permitió detectar diferencias significativas en los frames 68, 69, 73 y 80. El análisis de los mapas de atribución promedio por grupo (depresión y control), considerando únicamente estos pasos temporales, reveló una mayor atribución en los sujetos control, especialmente en regiones posteriores y centrales del cerebro. Esto sugiere que el modelo está capturando patrones característicos de los controles que están menos presentes en los sujetos con depresión. En

otras palabras, los resultados parecen reflejar una disminución funcional en regiones centrales y posteriores del cerebro en pacientes con depresión. Estos hallazgos se alinean con los resultados del estudio original de Bezmaternykh et al. (2021), que encontró una disminución de la conectividad funcional en la RMD, particularmente en la región posterior (precúneo y corteza cingulada posterior).

En conjunto, los resultados obtenidos muestran la alta complejidad en el desarrollo de modelos de DL aplicados a neuroimagen en contextos clínicos. A pesar de que no todas las estrategias implementadas ofrecieron mejoras cuantitativas, el estudio ha permitido identificar enfoques prometedores, validar decisiones metodológicas clave y establecer una base sólida sobre la que continuar el estudio de la depresión mediante modelos de DL.

5.2. Contribución a los ODS

Uno de los objetivos del presente proyecto fue tener en consideración varios Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030 de la ONU (Lee et al. 2016).

5.2.1. Sostenibilidad

El proyecto se alineó con el ODS 9 (Industria, innovación e infraestructura) y el ODS 12 (Producción y consumo responsables), ya que propuso un uso eficiente de los recursos tecnológicos y de datos abiertos. Más específicamente, se utilizaron datos de RMf previamente obtenidos. Esta práctica disminuye drásticamente las emisiones de CO_2 ya que como indican McAlister et al. (2022), una única sesión de RMf de menos de media hora puede generar aproximadamente 17 kg de CO_2. Asimismo, se emplearon modelos preentrenados en imágenes, evitando así entrenamientos desde cero con alto coste computacional y energético.

En la Tabla 5.1 se pueden observar las emisiones estimadas de CO_2 generadas durante el desarrollo de los modelos del presente proyecto. Nótese que se ha considerado únicamente el consumo energético asociado a las ejecuciones en el entorno de Google Colab, sin incluir el uso de equipos personales.

En conjunto, las decisiones tomadas en este proyecto reflejan un compromiso con una investigación más sostenible y eficiente. Aunque las emisiones totales han sido bajas, el uso de datos abiertos y modelos preentrenados demuestra que es posible avanzar en investigación sin recurrir a entrenamientos excesivamente costosos en términos ambientales.

Actividad	Emisiones estimadas (kg CO_2)
Exploración y preprocesamiento de datos RMf	0,050
Entrenamiento de modelos	0,745
Búsqueda de hiperparámetros con Optuna	0,151
Total estimado del proyecto	**0,946**

Tabla 5.1: Estimación de emisiones de CO_2 generadas durante el desarrollo de los modelos.

5.2.2. Comportamiento ético y responsabilidad social

El proyecto también se alineó con el ODS 3 (Salud y bienestar), al proponer herramientas innovadoras para mejorar el diagnóstico en salud mental, y con el ODS 16 (Paz, justicia e instituciones sólidas), al fomentar prácticas diagnósticas más transparentes y equitativas.

En cuanto al ODS 3, se desarrolló un modelo con alta sensibilidad (81,25 %) que, además, parecía detectar diferencias estructurales y funcionales acordes con la literatura especializada. Es decir, se contribuyó al avance del conocimiento, diagnóstico y tratamiento de la depresión. Respecto al ODS 16, este trabajo incluyó los resultados de todos los modelos desarrollados, independientemente de su rendimiento, así como una sección específica dedicada a las estrategias y técnicas que no funcionaron. Esto refuerza el compromiso con la transparencia y la reproducibilidad científica.

5.2.3. Diversidad y derechos humanos

El proyecto también se alineó con el ODS 5 (Igualdad de género) y el ODS 10 (Reducción de las desigualdades) al abordar los sesgos existentes en los diagnósticos clínicos.

A diferencia de los ODS anteriores, no se logró cumplir completamente con los objetivos del ODS 5 y 10. Esto se debe a que, por el tamaño reducido de la muestra, no fue posible la exclusión de muestras y en los modelos que aplicaron *data augmentation*, el equilibrio de clases se realizó en función del diagnóstico. Aun así, se resalta la importancia de este aspecto para el desarrollo de modelos clínicos válidos y justos.

5.3. Implicaciones y aportaciones

Los resultados obtenidos otorgan una serie de aportaciones al campo de los modelos de DL aplicados a la depresión.

En primer lugar, que el modelo haya obtenido unas métricas significativamente por encima del azar (50 %) sugiere que a pesar del escepticismo, las personas con depresión sí parecen presentar alteraciones estructurales y/o funcionales en el cerebro. Este resultado contribuye a un mayor entendimiento de este trastorno, con el objetivo de que en el futuro pueda desarrollarse un tratamiento más eficaz.

Los resultados de este proyecto también demuestran que el desarrollo de modelos de DL para la clasificación de la depresión es posible, incluso en condiciones limitadas como las que se han dado en este proyecto. De hecho, el modelo final alcanzó una sensibilidad del 81,25 %, una precisión superior a la de muchas herramientas actualmente utilizadas en el entorno clínico para el diagnóstico de trastornos psicológicos (Ayano et al. 2021). Por ello, se considera que esta línea de investigación merece seguir siendo explorada.

También se ha puesto de manifiesto el impacto que tienen las estrategias propuestas por Smucny, Shi & Davidson (2022). En concreto, este proyecto confirma que técnicas como el *data augmentation* y el *transfer learning* son esenciales para obtener modelos robustos cuando se trabaja con conjuntos de datos reducidos, como es habitual en estudios de neuroimagen debido al alto coste de la resonancia magnética funcional. Asimismo, se ha demostrado que las técnicas XAI son especialmente relevantes en contextos clínicos como este, ya que permiten verificar si los aspectos identificados por el modelo tienen coherencia con los hallazgos presentes en la literatura.

Por último, este estudio también aporta en términos de transparencia. Se expone no solo lo que ha funcionado, sino también lo que no ha dado buenos resultados bajo las condiciones de este proyecto, con el objetivo de que pueda servir de referencia para futuros investigadores que trabajen en entornos similares. Tal como señalan Janssen et al. (2018), la publicación de resultados negativos es clave para avanzar realmente en el desarrollo de modelos aplicables a la práctica clínica. En este sentido, este trabajo ofrece una metodología reproducible, con indicaciones claras sobre qué técnicas son viables y útiles en función de los datos, el entorno y la capacidad computacional disponible.

5.4. Limitaciones del proyecto

Como se ha mencionado anteriormente, la principal limitación del presente proyecto ha sido la capacidad computacional disponible. Esto ha condicionado decisiones clave a lo largo del

trabajo, como el uso de modelos en 2D en lugar de arquitecturas 3D, lo que ha supuesto una pérdida evidente de información espacial. Además, estas restricciones han impedido implementar técnicas y estructuras más complejas, como *co-teaching* o validación cruzada, que podrían haber mejorado la robustez y la generalización del modelo.

Otra limitación importante ha sido el reducido tamaño de la muestra. Los datos de RMf en abierto son escasos, especialmente los de depresión. Este aspecto ha limitado la capacidad de los modelos para aprender representaciones generales del problema y ha favorecido fenómenos de sobreajuste, especialmente en arquitecturas más profundas.

A nivel bibliográfico, se ha identificado una carencia notable de literatura centrada en la aplicación de DL a la depresión mediante datos de resonancia magnética funcional. Si bien existen estudios enfocados en otras patologías neurológicas, como el Alzheimer o el autismo (Chouliaras & O'Brien 2023), la investigación aplicada específicamente a depresión es todavía muy limitada. En particular, no se han encontrado trabajos previos que empleen arquitecturas híbridas basadas en CNN y RNN, lo que ha dificultado establecer comparaciones directas con el estado del arte y evaluar el impacto real del modelo desarrollado.

Adicionalmente, es importante destacar la desigualdad en la proporción de la muestra utilizada respecto al sexo biológico. En el grupo de pacientes con depresión, 38 participantes pertenecían al sexo biológico femenino mientras que 13 lo hacían con el masculino. Esta desigualdad también estaba presente en el grupo de control. Este aspecto es de gran relevancia, ya que se han documentado diferencias estructurales y funcionales en el cerebro en función del sexo biológico (DeCasien et al. 2022), por lo que los resultados obtenidos deben interpretarse con cautela, ya que podrían estar influenciados por un sesgo de género. En el presente estudio, debido al reducido tamaño muestral, no ha sido posible eliminar muestras para equilibrar por sexo, ya que el aumento de datos mediante data augmentation se ha aplicado para balancear por clase.

5.5. Líneas de trabajo futuro

A partir de los resultados obtenidos y de las limitaciones presentes en el proyecto, se proponen varias recomendaciones para investigaciones futuras que deseen continuar esta línea de trabajo.

En primer lugar, se recomienda explorar el uso de modelos tridimensionales cuando se disponga de recursos computacionales suficientes, ya que permitirían aprovechar de forma más completa la información espacial contenida en los datos de RMf. Asimismo, si los recursos computacionales lo permiten, sería interesante aplicar técnicas más complejas y avanzadas como validación cruzada o *co-teaching*.

A nivel de datos, una ampliación del tamaño muestral permitiría el desarrollo de modelos más robustos y generalizables. Sin embargo, como se ha mencionado anteriormente, existe una escasez de datos de RMf en abierto, especialmente en estudios centrados en pacientes con depresión. Por lo tanto, futuros proyectos deberían contemplar la posibilidad de solicitar el acceso a los conjuntos de datos originales utilizados en la literatura o, alternativamente, generar sus propios datos a través de colaboraciones clínicas o iniciativas de recolección propias.

Asimismo, se recomienda que los futuros estudios tengan en cuenta el sexo biológico como una variable crítica en el diseño experimental y el análisis de resultados. Dado que existen diferencias estructurales y funcionales documentadas entre cerebros de personas pertenecientes a diferente sexo biológico (DeCasien et al. 2022), es importante equilibrar las muestras en función del sexo. Esto evitaría posibles sesgos de género en la interpretación de los resultados y permitiría explorar si los patrones aprendidos por los modelos varían entre grupos.

Desde una perspectiva más amplia, este estudio ha mostrado la necesidad de fomentar la investigación específica en depresión mediante técnicas de DL, dado que la mayoría de trabajos existentes se centran en otras patologías neurológicas. Se destaca especialmente el interés de seguir explorando arquitecturas híbridas como CNN+RNN, ya que a pesar de no haber encontrado estudios previos con este enfoque en depresión, se considera prometedor. Además, se insiste en la importancia de compartir tanto los resultados positivos como los negativos, contribuyendo así a una ciencia más transparente y reproducible.

Finalmente, se recomienda seguir profundizando en el uso de técnicas de interpretabilidad y vincular sus resultados con hallazgos clínicos y neurobiológicos, para que estos modelos puedan convertirse en herramientas útiles de apoyo al diagnóstico en el futuro.

Bibliografía

Abrol, A., Fu, Z., Salman, M., Silva, R., Du, Y., Plis, S. & Calhoun, V. (2021), 'Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning', *Nature communications* **12**(1), 353.

Al Olaimat, M., Bozdag, S., Saeed, F., Initiative, A. D. N. et al. (2025), 'Ta-rnn: an attention-based time-aware recurrent neural network architecture to predict progression of alzheimer's disease', *Alzheimer's & Dementia* **20**(Suppl 1), e089010.

Ardalan, Z. & Subbian, V. (2022), 'Transfer learning approaches for neuroimaging analysis: a scoping review', *Frontiers in Artificial Intelligence* **5**, 780405.

Avberšek, L. K. & Repovš, G. (2022), 'Deep learning in neuroimaging data analysis: Applications, challenges, and solutions', *Frontiers in neuroimaging* **1**, 981642.

Avesta, A., Hossain, S., Lin, M., Aboian, M., Krumholz, H. M. & Aneja, S. (2023), 'Comparing 3d, 2.5 d, and 2d approaches to brain image auto-segmentation', *Bioengineering* **10**(2), 181.

Ayano, G., Demelash, S., Yohannes, Z., Haile, K., Tulu, M., Assefa, D., Tesfaye, A., Haile, K., Solomon, M., Chaka, A. et al. (2021), 'Misdiagnosis, detection rate, and associated factors of severe psychiatric disorders in specialized psychiatry centers in ethiopia', *Annals of general psychiatry* **20**, 1–10.

Ball, J. R., Miller, B. T. & Balogh, E. P. (2015), *Improving diagnosis in health care*, National Academies Press.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J. & Erbaugh, J. (1961), 'An inventory for measuring depression', *Archives of general psychiatry* **4**(6), 561–571.

Bernal, J., Kushibar, K., Asfaw, D. S., Valverde, S., Oliver, A., Martí, R. & Lladó, X. (2019), 'Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review', *Artificial intelligence in medicine* **95**, 64–81.

Bezmaternykh, D. D., Melnikov, M. Y., Savelov, A. A., Kozlova, L. I., Petrovskiy, E. D., Natarova, K. A. & Shtark, M. B. (2021), 'Brain networks connectivity in mild to moderate depression: resting state fmri study with implications to nonpharmacological treatment', *Neural plasticity* **2021**(1), 8846097.

Bhati, D., Neha, F. & Amiruzzaman, M. (2024), 'A survey on explainable artificial intelligence (xai) techniques for visualizing deep learning models in medical imaging', *Journal of Imaging* **10**(10), 239.

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L. et al. (2021), 'Hyperparameter optimization: Foundations, algorithms, best practices and open challenges. arxiv', *arXiv preprint arXiv:2107.05847* .

Bunge, S. A. & Kahn, I. (2009), 'Cognition: An overview of neuroimaging techniques', *Encyclopedia of Neuroscience* **2**, 1063–1067.

Castanheira, L., Silva, C., Cheniaux, E. & Telles-Correia, D. (2019), 'Neuroimaging correlates of depression—implications to clinical practice', *Frontiers in Psychiatry* **10**, 703.

Chau, W. & McIntosh, A. R. (2005), 'The talairach coordinate of a point in the mni space: how to interpret it', *Neuroimage* **25**(2), 408–416.

Chivite, A. (2025), 'Repositorio de código del proyecto: Deep learning para diagnostico de depresión en rmf', https://github.com/achivite/Deep_Learning_para_Diagnostico_de_Depresion_en_RMf. 28 de mayo de 2025.

Chollet, F. (2025), 'Guide to saving and serializing models with keras in tensorflow 2.0', https://colab.research.google.com/drive/172D4jishSgE3N7AO6U2OKAA_0wNnrMOq#scrollTo=mJqOnOsnzCRy. Google Colaboratory notebook.

Chouliaras, L. & O'Brien, J. T. (2023), 'The use of neuroimaging techniques in the early and differential diagnosis of dementia', *Molecular Psychiatry* **28**(10), 4084–4097.

Davatzikos, C. (2019), 'Machine learning in neuroimaging: Progress and challenges'.

DeCasien, A. R., Guma, E., Liu, S. & Raznahan, A. (2022), 'Sex differences in the human brain: a roadmap for more careful analysis and interpretation of a biological reality', *Biology of Sex Differences* **13**(1), 43.

Dufumier, B., Gori, P., Petiton, S., Louiset, R., Mangin, J.-F., Grigis, A. & Duchesnay, E. (2024), 'Exploring the potential of representation and transfer learning for anatomical neuroimaging: application to psychiatry', *NeuroImage* **296**, 120665.

Dunlop, B. W. & Mayberg, H. S. (2017), Neuroimaging advances for depression, *in* 'Cerebrum: the Dana forum on brain science', Vol. 2017, pp. cer–16.

Eitel, F., Schulz, M.-A., Seiler, M., Walter, H. & Ritter, K. (2021), 'Promises and pitfalls of deep neural networks in neuroimaging-based psychiatric research', *Experimental Neurology* **339**, 113608.

Farahani, F. V., Fiok, K., Lahijanian, B., Karwowski, W. & Douglas, P. K. (2022), 'Explainable ai: A review of applications to neuroimaging data', *Frontiers in Neuroscience* **16**, 906290.

First, M. B., Drevets, W. C., Carter, C., Dickstein, D. P., Kasoff, L., Kim, K. L., McConathy, J., Rauch, S., Saad, Z. S., Savitz, J. et al. (2018), 'Clinical applications of neuroimaging in psychiatric disorders', *American Journal of Psychiatry* **175**(9), 915–916.

Floyd, B. J. (1997), 'Problems in accurate medical diagnosis of depression in female patients', *Social science & medicine* **44**(3), 403–412.

García-Gutiérrez, M. S., Navarrete, F., Sala, F., Gasparyan, A., Austrich-Olivares, A. & Manzanares, J. (2020), 'Biomarkers in psychiatry: concept, definition, types and relevance to the clinical reality', *Frontiers in psychiatry* **11**, 432.

Glover, G. H. (2011), 'Overview of functional magnetic resonance imaging', *Neurosurgery Clinics of North America* **22**(2), 133.

Gray, J. P., Müller, V. I., Eickhoff, S. B. & Fox, P. T. (2020), 'Multimodal abnormalities of brain structure and function in major depressive disorder: a meta-analysis of neuroimaging studies', *American Journal of Psychiatry* **177**(5), 422–434.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I. & Sugiyama, M. (2018), Co-teaching: Robust training of deep neural networks with extremely noisy labels, *in* 'Advances in Neural Information Processing Systems', Vol. 31, Curran Associates, Inc.

Henderson, T. A., Van Lierop, M. J., McLean, M., Uszler, J. M., Thornton, J. F., Siow, Y.-H., Pavel, D. G., Cardaci, J. & Cohen, P. (2020), 'Functional neuroimaging in psychiatry—aiding in diagnosis and guiding treatment. what the american psychiatric association does not know', *Frontiers in psychiatry* **11**, 276.

Jaber, H. A., Aljobouri, H. K., Çankaya, İ., Koçak, O. M. & Algin, O. (2019), 'Preparing fmri data for postprocessing: Conversion modalities, preprocessing pipeline, and parametric and nonparametric approaches', *IEEE Access* **7**, 122864–122877.

Janssen, R. J., Mourão-Miranda, J. & Schnack, H. G. (2018), 'Making individual prognoses in psychiatry using neuroimaging and machine learning', *Biological psychiatry: Cognitive neuroscience and neuroimaging* **3**(9), 798–808.

Kalin, N. H. (2021), 'Understanding the value and limitations of mri neuroimaging in psychiatry', *American Journal of Psychiatry* **178**(8), 673–676.

Karimi, D., Dou, H., Warfield, S. K. & Gholipour, A. (2020), 'Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis', *Medical image analysis* **65**, 101759.

Kim, J. S. (2025), 'A novel approach for brain connectivity using recurrent neural networks and integrated gradients', *Computers in Biology and Medicine* **184**, 109404.

Krishnapriya, S. & Karuna, Y. (2023), 'Pre-trained deep learning models for brain mri image classification', *Frontiers in Human Neuroscience* **17**, 1150120.

Lee, B. X., Kjaerulf, F., Turner, S., Cohen, L., Donnelly, P. D., Muggah, R., Davis, R., Realini, A., Kieselbach, B., MacGregor, L. S. et al. (2016), 'Transforming our world: implementing the 2030 agenda through sustainable development goal indicators', *Journal of public health policy* **37**, 13–31.

Lee, K.-H., He, X., Zhang, L. & Yang, L. (2018), Cleannet: Transfer learning for scalable image classifier training with label noise, *in* '2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 5447–5456.

Lim, G. Y., Tam, W. W., Lu, Y., Ho, C. S., Zhang, M. W. & Ho, R. C. (2018), 'Prevalence of depression in the community from 30 countries between 1994 and 2014', *Scientific reports* **8**(1), 2861.

Lorenzo, S. M., Astier-Peña, M. P. & Benejam, T. C. (2021), 'El error diagnóstico y sobrediagnóstico en atención primaria. propuestas para la mejora de la práctica clínica en medicina de familia', *Atención Primaria* **53**, 102227.

Mao, Z., Su, Y., Xu, G., Wang, X., Huang, Y., Yue, W., Sun, L. & Xiong, N. (2019), 'Spatio-temporal deep learning method for adhd fmri classification', *Information Sciences* **499**, 1–11.

McAlister, S., McGain, F., Breth-Petersen, M., Story, D., Charlesworth, K., Ison, G. & Barratt, A. (2022), 'The carbon footprint of hospital diagnostic imaging in australia', *The Lancet Regional Health–Western Pacific* **24**.

McElroy, E., Fearon, P., Belsky, J., Fonagy, P. & Patalay, P. (2018), 'Networks of depression and anxiety symptoms across development', *Journal of the American Academy of Child & Adolescent Psychiatry* **57**(12), 964–973.

Mienye, I. D., Swart, T. G. & Obaido, G. (2024), 'Recurrent neural networks: A comprehensive review of architectures, variants, and applications', *Information* **15**(9), 517.

Montgomery, S. A. & Åsberg, M. (1979), 'A new depression scale designed to be sensitive to change', *The British journal of psychiatry* **134**(4), 382–389.

Moreno-Agostino, D., Wu, Y.-T., Daskalopoulou, C., Hasan, M. T., Huisman, M. & Prina, M. (2021), 'Global trends in the prevalence and incidence of depression: a systematic review and meta-analysis', *Journal of affective disorders* **281**, 235–243.

Mumuni, A. & Mumuni, F. (2022), 'Data augmentation: A comprehensive survey of modern approaches', *Array* **16**, 100258.

Najafpour, Z., Fatemi, A., Goudarzi, Z., Goudarzi, R., Shayanfard, K. & Noorizadeh, F. (2021), 'Cost-effectiveness of neuroimaging technologies in management of psychiatric and insomnia disorders: A meta-analysis and prospective cost analysis', *Journal of Neuroradiology* **48**(5), 348–358.

Nour, M. M., Liu, Y. & Dolan, R. J. (2022), 'Functional neuroimaging in psychiatry and the case for failing better', *Neuron* **110**(16), 2524–2544.

Prvulovic, D. & Hampel, H. (2010), '¿ un cambio de paradigma en el diagnóstico psiquiátrico moderno?: Anomalías en las redes neuronales como concepto fisiopatológico y nueva herramienta de diagnóstico', *Revista de psiquiatría y salud mental* **3**(4), 115–118.

Qin, K., Lei, D., Pinaya, W., Pan, N., Li, W., Zhu, Z., Sweeney, J., Mechelli, A. & Gong, Q. (2022), 'Using graph convolutional network to characterize individuals with major depressive disorder across multiple imaging sites', *EBioMedicine* **78**, 103977.

Quaak, M., van de Mortel, L., Thomas, R. M. & van Wingen, G. (2021), 'Deep learning applications for the classification of psychiatric disorders using neuroimaging data: systematic review and meta-analysis', *NeuroImage: Clinical* **30**, 102584.

Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N. & Fanos, V. (2020), 'Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment', *Medicina* **56**(9), 455.

Ramu, A. & Haldorai, A. (2024), 'Techniques advantages and limitations of neuroimaging: A systematic review', *Journal of Biomedical and Sustainable Healthcare Applications* **54**, 62.

Ritt, P. (2022), 'Recent developments in spect/ct', *Seminars in Nuclear Medicine* **52**(3), 276–285.

Sánchez Fernández, I. & Peters, J. M. (2023), 'Machine learning and deep learning in medicine and neuroimaging', *Annals of the Child Neurology Society* **1**(2), 102–122.

Sarracanie, M., LaPierre, C. D., Salameh, N., Waddington, D. E., Witzel, T. & Rosen, M. S. (2015), 'Low-cost high-performance mri', *Scientific reports* **5**(1), 15177.

Schiff, G. D., Hasan, O., Kim, S., Abrams, R., Cosby, K., Lambert, B. L., Elstein, A. S., Hasler, S., Kabongo, M. L., Krosnjar, N. et al. (2009), 'Diagnostic error in medicine: analysis of 583 physician-reported errors', *Archives of internal medicine* **169**(20), 1881–1887.

Schmaal, L., Pozzi, E., C. Ho, T., Van Velzen, L. S., Veer, I. M., Opel, N., Van Someren, E. J., Han, L. K., Aftanas, L., Aleman, A. et al. (2020), 'Enigma mdd: seven years of global neuroimaging studies of major depression through worldwide data sharing', *Translational psychiatry* **10**(1), 172.

Signorile, W. J., Mahajan, A., Fulbright, R. K. & Zubair, A. S. (2024), 'Comparative analysis of energy expenditure and costs in neuroimaging', *Journal of the Neurological Sciences* **460**, 123001.

Smucny, J., Shi, G. & Davidson, I. (2022), 'Deep learning in neuroimaging: overcoming challenges with emerging approaches', *Frontiers in Psychiatry* **13**, 912600.

Smucny, J., Shi, G., Lesh, T. A., Carter, C. S. & Davidson, I. (2022), 'Data augmentation with mixup: Enhancing performance of a functional neuroimaging-based prognostic deep learning classifier in recent onset psychosis', *NeuroImage: Clinical* **36**, 103214.

Solovyev, R., Kalinin, A. A. & Gabruseva, T. (2022), '3d convolutional neural networks for stalled brain capillary detection', *Computers in biology and medicine* **141**, 105089.

Song, H., Kim, M., Park, D., Shin, Y. & Lee, J.-G. (2022), 'Learning from noisy labels with deep neural networks: A survey', *IEEE transactions on neural networks and learning systems* **34**(11), 8135–8153.

Sujatha, C. et al. (2021), Identification of schizophrenia using lstm recurrent neural network, *in* '2021 seventh international conference on bio signals, images, and instrumentation (ICBSII)', IEEE, pp. 1–6.

Sundararajan, M., Taly, A. & Yan, Q. (2017), Axiomatic attribution for deep networks, *in* 'International conference on machine learning', PMLR, pp. 3319–3328.

Teplin, L. A., Abram, K. M. & Luna, M. J. (2023), 'Racial and ethnic biases and psychiatric misdiagnoses: toward more equitable diagnosis and treatment'.

Thomas, A. W., Heekeren, H. R., Müller, K.-R. & Samek, W. (2019), 'Analyzing neuroimaging data through recurrent deep learning models', *Frontiers in neuroscience* **13**, 1321.

Van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G. & Viergever, M. A. (2022), 'Explainable artificial intelligence (xai) in deep learning-based medical image analysis', *Medical Image Analysis* **79**, 102470.

Varshney, R. K., Katiyar, A. & Johri, P. (2025), Hybrid cnn-rnn models for multimodal analysis of autism spectrum disorder neuroimaging, *in* '2025 International Conference on Automation and Computation (AUTOCOM)', IEEE, pp. 155–160.

Vermani, M., Marcus, M. & Katzman, M. A. (2011), 'Rates of detection of mood and anxiety disorders in primary care: a descriptive, cross-sectional study', *The primary care companion for CNS disorders* **13**(2), 27211.

Wald, L. L., McDaniel, P. C., Witzel, T., Stockmann, J. P. & Cooley, C. Z. (2020), 'Low-cost and portable mri', *Journal of Magnetic Resonance Imaging* **52**(3), 686–696.

World Health Organization (2017), *Depression and other common mental disorders: global health estimates*, World Health Organization.

Yan, C.-G., Chen, X., Li, L., Castellanos, F. X., Bai, T.-J., Bo, Q.-J., Cao, J., Chen, G.-M., Chen, N.-X., Chen, W. et al. (2019), 'Reduced default mode network functional connectivity in patients with recurrent major depressive disorder', *Proceedings of the National Academy of Sciences* **116**(18), 9078–9083.

Yan, W., Qu, G., Hu, W., Abrol, A., Cai, B., Qiao, C., Plis, S. M., Wang, Y.-P., Sui, J. & Calhoun, V. D. (2022), 'Deep learning in neuroimaging: Promises and challenges', *IEEE Signal Processing Magazine* **39**(2), 87–98.

Yang, L. & Shami, A. (2020), 'On hyperparameter optimization of machine learning algorithms: Theory and practice', *Neurocomputing* **415**, 295–316.

Yen, C., Lin, C.-L. & Chiang, M.-C. (2023), 'Exploring the frontiers of neuroimaging: a review of recent advances in understanding brain functioning and disorders', *Life* **13**(7), 1472.

Yu, T. & Zhu, H. (2020), 'Hyper-parameter optimization: A review of algorithms and applications', *arXiv preprint arXiv:2003.05689* .

Zhao, Z., Chuah, J. H., Lai, K. W., Chow, C.-O., Gochoo, M., Dhanalakshmi, S., Wang, N., Bao, W. & Wu, X. (2023), 'Conventional machine learning and deep learning in alzheimer's disease diagnosis using neuroimaging: A review', *Frontiers in computational neuroscience* **17**, 1038636.