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, 15 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):	15 de mayo de 2025
Titulación o programa:	Máster Universitario en Ciencia de Datos (Data Science)
Área del Trabajo Final:	5
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 incidencia está en aumento. El diagnóstico se basa en evaluaciones clínicas que presentan un porcentaje de error considerable. En este contexto, la neuroimagen, junto con la aplicación de modelos de aprendizaje automático, ha emergido como una herramienta prometedora para el estudio de los trastornos mentales. Sin embargo, su aplicación sigue siendo limitada debido a la falta de modelos computacionales robustos.

El presente proyecto propone el desarrollo de un modelo de *deep learning* híbrido que combine redes neuronales convolucionales y recurrentes con la finalidad de clasificar correctamente imágenes de resonancias magnéticas funcionales en reposo de personas deprimidas y personas que no lo están. Para mejorar el rendimiento y la interpretabilidad del modelo, se aplicarán técnicas del estado del arte como *transfer learning*, *data augmentation*, y *explainable artificial intelligence*. Además, dado que cierto número de diagnósticos pueden ser erróneos, y por lo tanto, de las etiquetas en los datos, se emplearán métodos de reducción de ruido en las etiquetas.

El objetivo del proyecto es obtener un modelo con una alta sensibilidad y capacidad de generalización que pueda asistir en el diagnóstico de la depresión y reducir los diagnósticos erróneos.

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 incidence is increasing. Diagnosis is based on clinical assessments that have a considerable error rate. In this context, neuroimaging, together with the application of machine learning models, has emerged as a promising tool for the study of mental disorders. However, its application remains limited due to the lack of robust computational 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 MRI images of depressed and non-depressed individuals. To improve the performance and interpretability of the model, state-of-the-art techniques such as transfer learning, data augmentation, and explainable artificial intelligence will be applied. In addition, since a certain number of diagnoses may be erroneous, and therefore, of the labels in the data, label noise reduction methods will be employed.

The goal of the project is to obtain a model with high sensitivity and generalizability that can assist in the diagnosis of depression and reduce misdiagnosis.

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

Índice general

Índice de figuras

Índice de cuadros

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 deprimidas en comparación con personas sanas (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 diagnóstico 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 los pacientes, 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 aprendizaje automático capaz de identificar si un paciente 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 estudio 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 espacial y 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 sufre de diagnósticos clínicos imprecisos, se aplicarán diferentes técnicas de tratamiento del ruido en las etiquetas.

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 depresión y de personas sanas. 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 indica la activación cerebral en diferentes intensidades, para lo cual las CNN son ideales. Por el otro lado, la información temporal indica 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 pacientes con 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. 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 alinea 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 identificadas con género 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. Además, al final de cada fase se requerirá la elaboración de distintos entregables, con el objetivo de documentar los avances obtenidos en cada etapa.

- **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 el código implementado.
- **M4 – Redacción de la documentación:** elaboración de la memoria final del proyecto y de una presentación audiovisual breve.
- **M5 – Defensa del proyecto:** entrega de la documentación del proyecto y exposición oral del trabajo realizado ante el tribunal académico.

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

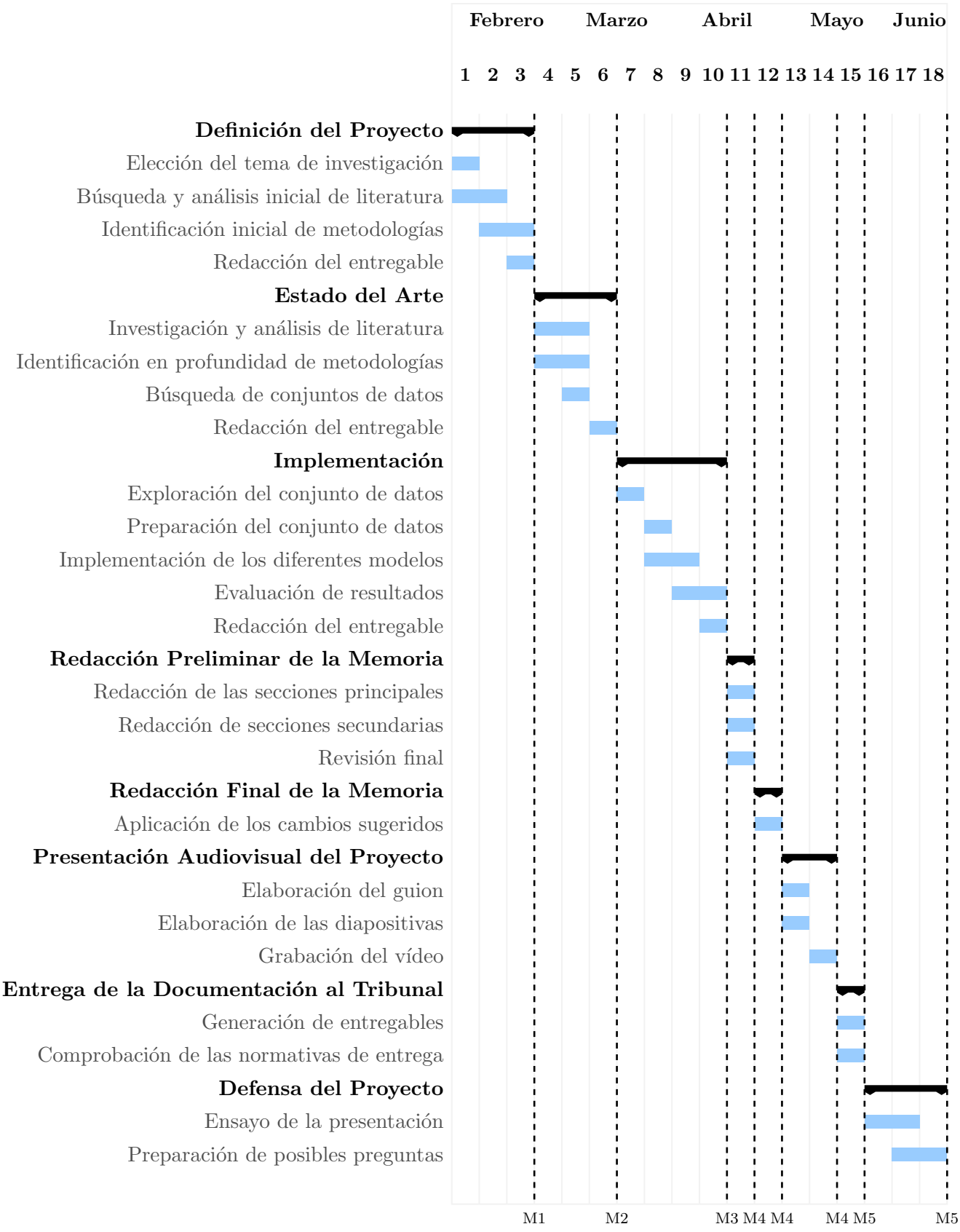

Figura 1.1: Planificación temporal del proyecto.

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 etiquetas potencialmente erróneas.

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 siguiente tabla 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.

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

Cuadro 2.1: Biomarcadores asociados con la depresión.

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 aprendizaje profundo representa una oportunidad para descubrir patrones latentes que no son fácilmente identificables mediante métodos estadísticos tradicionales.

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).

Cada técnica de neuroimagen realiza distintas mediciones del cerebro, lo que se traduce en diferencias tanto en funcionales como operativas. Algunas 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 y el impacto medioambiental (Najafpour et al. 2021, Signorile et al. 2024). En la siguiente tabla se resumen algunas de estas características clave de algunas de las técnicas más relevantes.

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

Cuadro 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 o sin trastornos psiquiátricos como la depresión. Además, en los últimos años se están logrando avances significativos en el desarrollo de escáneres más potentes y precisos (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 *Deep Learning (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 el siguiente gráfico, 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.

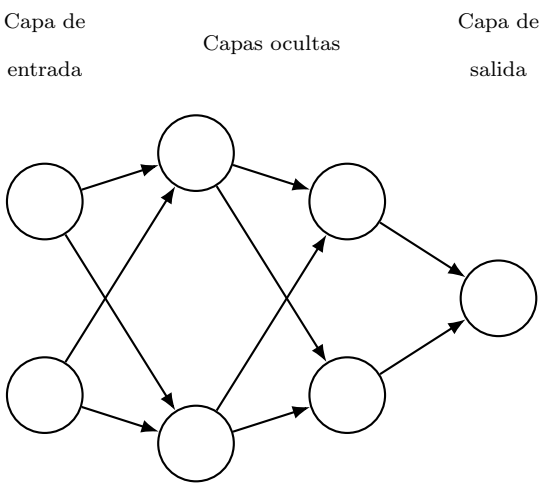

Figura 2.1: Arquitectura básica de una ANN.

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 sor sus siglas en inglés) y las redes neuronales recurrentes (RNN por sus siglas en ingles).

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 dimen-

sionalidad 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 el siguiente gráfico.

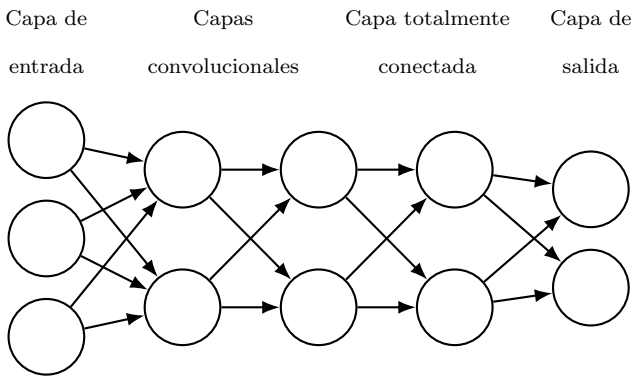

Figura 2.2: Arquitectura básica de una CNN.

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 depresión mayor y sujetos sanos. 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 el siguiente gráfico.

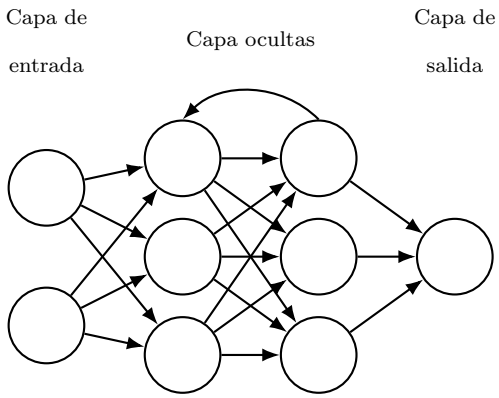

Figura 2.3: Arquitectura básica de una RNN.

Por ejemplo, Sujatha et al. (2021) desarrollaron una RNN utilizando datos de RMf en reposo

para clasificar pacientes con esquizofrenia y sujetos sanos, 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 CNN 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 pacientes con trastorno por déficit de atención e hiperactividad (TDAH) y sujetos sanos. En su enfoque, las CNN 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 como *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 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)

La XAI es un conjunto de técnicas que permiten que los modelos de DL no solo proporcionen métricas de precisión, sino también información sobre qué variables utilizaron durante el entrenamiento y la toma de decisiones. Entre las técnicas de XAI se incluyen los *saliency mapping methods* que proporcionan explicaciones a nivel de instancia individual y los *Signal*

Reconstruction Methods que proporcionan explicaciones a nivel de características de salida, entre otras (Qian et al. 2023).

Por ejemplo, Chavez et al. (2025) desarrollaron un modelo explicable basado en *Random Forest* y *Shapley Additive Explanations (SHAP)* para clasificar esquizofrenia a partir de datos de RM. El modelo alcanzó una precisión del 72 % e identificó regiones clave relacionadas con la esquizofrenia como la corteza fusiforme temporal superior.

En conclusión, estas técnicas no solo hacen que los modelos de DL sean mas potentes, sino que también aumentan su fiabilidad, 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 las etiquetas erróneas

Otra importante limitación que los modelos de DL presentan es su sensibilidad a las etiquetas erróneas. En el contexto de la neuroimagen, esto puede implicar por ejemplo que los datos de un paciente diagnosticado 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 de 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 las etiquetas erróneas 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 las etiquetas erróneas.
- 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 etiquetas erroneas.
- Procedimientos de entrenamiento: uso de procedimientos de entrenamiento alternativos para evitar el sobreajuste a las etiquetas erróneas.

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 etiquetas erróneas 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 etiquetas erróneas. Mediante esta estrategia, lograron reducir en un 41,5 % el número de etiquetas erróneas en clases no verificadas, lo que demuestra su efectividad con muy poca supervisión.

En conclusión, existen diversas estrategias para gestionar etiquetas erróneas 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 llega nuestro conocimiento, no ha sido aplicada previamente en el estudio de los biomarcadores de la depresión mediante datos de RMf.

Capítulo 3

Materiales y métodos

En este capítulo se describen los recursos utilizados y la metodología implementada para el desarrollo del proyecto. Más específicamente, se detallan las herramientas utilizadas, las características del conjunto de datos empleado, las técnicas de preprocesamiento aplicadas a las imágenes de RMf, los diferentes modelos desarrollados y las técnicas empleadas para su entrenamiento, ajuste e interpretación.

3.1. Fuente y descripción del conjunto de datos

El conjunto de datos empleado proviene del estudio de Bezmaternykh et al. (2021). Los datos se tratan de 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 mm, tiempo de repetición = 2500ms, tiempo de eco = 35 ms, 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 3.1 se puede observar el corte central axial de uno de los participantes en el estudio.

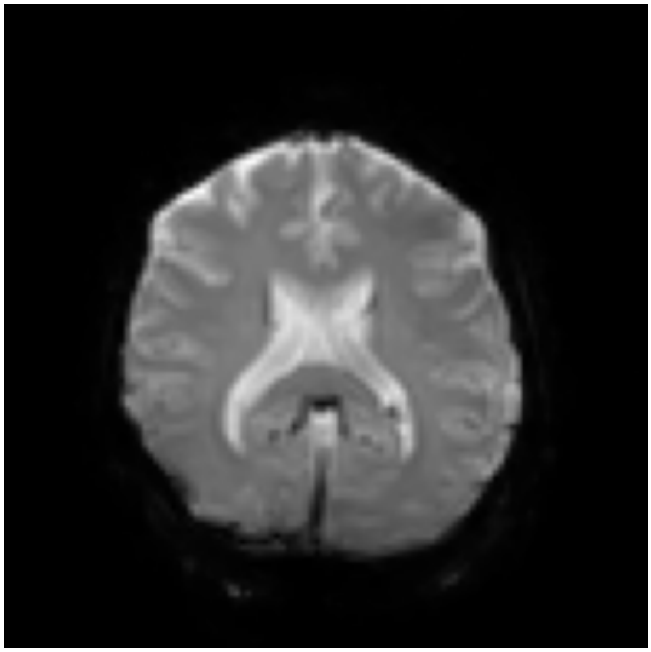

Figura 3.1: Corte axial central de un participante.

La muestra está compuesta por 72 participantes, incluyendo pacientes diagnosticados con depresión leve o moderada y controles sanos. Entre las variables clínicas utilizadas se encuentran varias escalas psicométricas utilizadas para evaluar la sintomatología depresiva y otros factores psicológicos relevantes como la *Montgomery–Asberg Depression Rating Scale* o el *Beck Depression Inventory*, entre otros. Estas métricas permiten caracterizar la severidad del cuadro clínico y complementar la interpretación de los datos neurofuncionales, aunque no se utilizaron directamente en los modelos desarrollados. El rango de edad media de los participantes es de 33.1 ± 9.5 años en los pacientes diagnosticados con depresión y de 33.8 ± 8.5 en los controles sanos. La distribución de los participantes por género y grupo diagnóstico se representa en la siguiente figura.

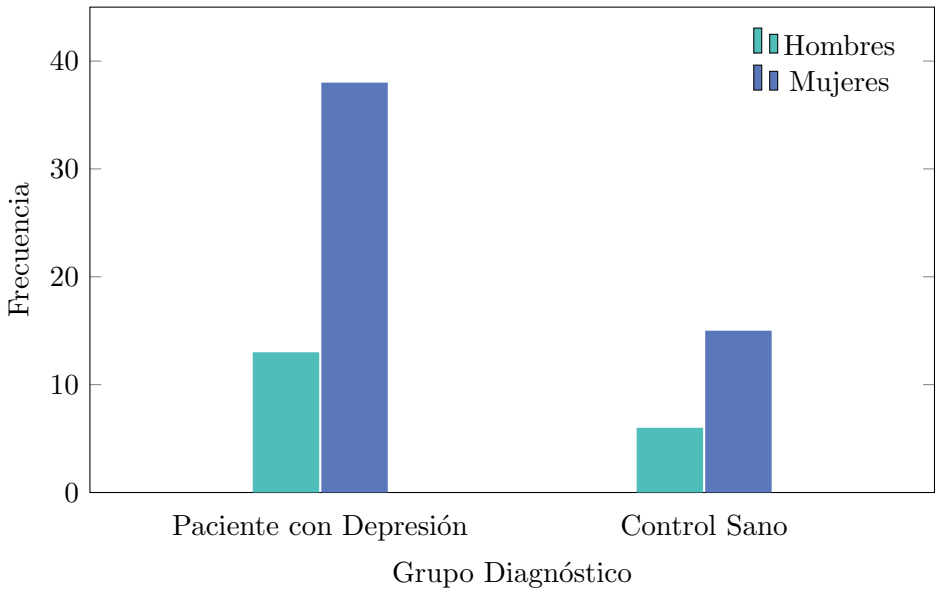

Figura 3.2: Distribución de participantes por grupo diagnóstico y género.

3.2. Entorno computacional y herramientas utilizadas

La implementación del proyecto se ha realizado 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, se ha hecho uso de las siguientes librerías:

- *nilearn.masking* y *nibabel*, utilizadas para la carga, visualización y enmascaramiento de los datos de resonancia magnética funcional (RMf) en formato NIfTI.
- *tensorflow.keras.layers*, para la definición de arquitecturas de CNN y RNN.
- *tensorflow.keras.applications*, para la incorporación de modelos preentrenados mediante técnicas de *transfer learning*.
- *tensorflow.keras.preprocessing*, empleada en las tareas de *data augmentation*.
- *tensorflow.keras.losses*, utilizada para la implementación de técnicas de manejo de etiquetas ruidosas.
- *tf-explain*, utilizada para la generación de mapas de activación mediante Grad-CAM como técnica XAI.
- *optuna*, empleada para la optimización de hiperparámetros.

Estas librerías representan 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.

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	*tensorflow.keras.preprocessing*
Manejo de etiquetas ruidosas	*tensorflow.keras.losses*
Explicabilidad del modelo (XAI - GradCAM)	*tf-explain*
Ajuste fino de hiperparámetros	*optuna*

Cuadro 3.1: Librerías utilizadas durante el desarrollo de las diferentes fases del proyecto.

El desarrollo se ha llevado a cabo en el entorno de ejecución *Google Colaboratory*, mediante notebooks diferenciados para cada fase del proyecto. Los archivos de datos y resultados han sido 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 (splits) debido a su tamaño, se encuentra disponible en el siguiente repositorio público de GitHub: *https://github.com/achivite/Deep_Learning_para_Diagnostico_de_Depresion_en_RMf*.

3.3. Estrategia metodológica y diseño experimental

La metodología del presente proyecto se fundamenta en una serie de estudios clave que abordan los desafíos actuales en el uso de DL para el análisis de datos de RMf.

Como se ha mencionado anteriormente, los datos de resonancia magnética funcional (RMf) contienen información tanto espacial como temporal. Por esta razón, se ha adoptado una arquitectura híbrida CNN (para el análisis espacial) y RNN (para capturar la dinámica temporal). Concretamente, se ha empleado 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), han identificado 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) proponen una serie de estrategias: el uso de *transfer learning* y *data augmentation* para contrarrestar la escasez de datos, así como técnicas de inteligencia artificial explicable (XAI) para mejorar la interpretabilidad de los modelos.

La estrategia de *transfer learning* adoptada en este estudio sigue la propuesta de Krishnapriya & Karuna (2023), quienes implementan distintos modelos CNN preentrenados sobre bases de datos de imágenes de RM. Los modelos utilizados en su estudio incluyen VGG-16, VGG-19, ResNet50 e Inception V3. En el presente trabajo se han replicado 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 ha optado por congelar todas las capas convolucionales y reentrenar únicamente la capa completamente conectada final, siguiendo también el procedimiento descrito en el estudio mencionado.

La estrategia de *data augmentation* implementada también se basa en el enfoque de Krishnapriya & Karuna (2023), que 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.

Respecto a la interpretabilidad de los modelos, se ha implementado Grad-CAM con el objetivo de identificar las regiones anatómicas relevantes dentro de cada corte axial, permitiendo así una mejor comprensión de los patrones espaciales que el modelo utiliza para la toma de decisiones.

Con el objetivo de abordar la posible presencia de etiquetas erróneas en los datos, se han evaluado distintas funciones de pérdida robustas, de acuerdo con las recomendaciones de Karimi et al. (2020). Asimismo, se ha realizado una optimización de hiperparámetros mediante búsqueda automatizada, una etapa fundamental en el diseño y evaluación de modelos de ML (Bischl et al. 2021).

En resumen, la metodología desarrollada en este trabajo consiste en la aplicación incremental y combinada de estas estrategias sobre un modelo CNN+GRU entrenado con datos de RMf de sujetos con y sin diagnóstico de depresión. La siguiente sección detalla la estructura concreta

de esta metodología y su implementación progresiva.

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+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	Optimización de hiperparámetros	Afinar el rendimiento del modelo más prometedor	Ajuste automático con *Optuna*
8	Modelos con manejo de etiquetas ruidosas	Incrementar la robustez ante errores en las etiquetas	*Focal Loss*, *iMAE*, *Label Smoothing*
9	Técnicas de interpretabilidad	Analizar las regiones y momentos relevantes para el modelo	*Grad-CAM*, *Occlusion temporal*

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

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 a espacio MNI con voxelado de 2x2x2 mm.
- Suavizado con un kernel gaussiano de 8 mm FWHM.

En el presente proyecto se ha optado 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 implementarán en la fase de *transfer learning* requieren que sus datos de entrada estén 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 adoptado en este estudio se basa en el análisis secuencial de un único corte axial por participante (el corte con mayor varianza) no se aplicará 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 aplicará suavizado espacial ya que el suavizado tridimensional implicaría 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 fase de exploración de datos, las tareas de preprocesado que se llevarán a cabo en el 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 almacenan en formato *.npy*, propio de *NumPy*, con el objetivo de optimizar tanto el tiempo de carga como el rendimiento durante el entrenamiento de los modelos. Esta estrategia permite separar el *pipeline* de preprocesamiento de la etapa de modelado y facilita la reproducibilidad de los experimentos.

Además del preprocesamiento general, se han definido dos fases adicionales de preparación de datos, adaptadas a los requisitos específicos de las secciones de *data augmentation* y *transfer learning*. La metodología aplicada en ambas secciones sigue las directrices del estudio de Krishnapriya & Karuna (2023), en el que se evalúan diversos modelos CNN preentrenados aplicados a imágenes de resonancia magnética. En dicho estudio, se aplican las siguientes transformaciones para la aumentación artificial del conjunto de datos: rotación del 15 %, desplazamiento horizontal y vertical del 5 %, escalado mediante un factor de 1/255 y duplicación en los ejes horizontal y vertical. Con el fin de mantener la coherencia metodológica, en el presente trabajo se replican dichas transformaciones, con la única excepción del re-escalado, que ha sido omitido. Adicionalmente, con el fin de equilibrar el conjunto de datos original, se aplicó el *data augmentation* con una proporción de $\times 5$ para la clase control y $\times 2$ para la clase depresiva.

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

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 estudio comparten una serie de hiperparámetros base comunes, que fueron seleccionados por su eficacia y estabilidad durante las primeras fases de experimentación. La Tabla 4.1 resume estos valores, que sirvieron como configuración inicial para todos los modelos, salvo en aquellos casos en los que se indica explícitamente una modificación.

Hiperparámetro	**Valor**
Regularización L2	1e-4
Optimizador	Adam
Tasa de aprendizaje	0.0001
Unidades GRU	64
Unidades en capa densa	64
Dropout convolucional	0.2
Dropout final	0.5
Activación final	Sigmoid
Métricas	Accuracy, Recall, Precision, AUC
Número de épocas	30

Cuadro 4.1: Hiperparámetros base comunes a los modelos implementados.

4.2. Resultados de modelos base

Esta sección tenía un doble objetivo: sentar una base sobre la cual comparar el resultado de la aplicación de las técnicas de aumento de robustez de los modelos, y examinar si la inclusión de la dimensión temporal mediante la RNN produciría mejores resultados.

La Tabla 4.2 recoge los resultados obtenidos por estos modelos base implementados sin ninguna técnica adicional. Ambos modelos lograron una sensibilidad perfecta (100 %), aunque con una especificidad nula (0 %), lo que sugiere un sesgo completo hacia la predicción positiva. En comparación, el modelo CNN+GRU presentó un área bajo la curva (AUC) ligeramente superior y un menor coste computacional en tiempo y emisiones.

Métrica	**CNN**	**CNN+GRU**
Exactitud	0.7273	0.7273
Precisión	0.7273	0.7273
Sensibilidad	1.0000	1.0000
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

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

Los resultados cumplen con el primer objetivo de sentar una base sobre la cual comparar los futuros modelos, pero no han permitido observar una mejora significativa con la inclusión de la RNN.

4.3. Resultados de *data augmentation*

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

La Tabla 4.3 presenta los resultados del mismo modelo CNN+GRU anterior, pero entrenado con una muestra mayor. A pesar de lograr una sensibilidad perfecta (100 %), al igual que el modelo base CNN+GRU, el resto de métricas muestran un descenso, especialmente en precisión (0.5000) y F1-score (0.6667). El área bajo la curva (AUC) también fue inferior (0.4062 frente a 0.4167 en el modelo base), lo que indica que la estrategia de aumentación no mejoró el

rendimiento general del modelo. Además, tanto el tiempo de entrenamiento como las emisiones de CO_2 se mantuvieron similares.

Métrica	**CNN+GRU + Augmentation**
Exactitud	0.5000
Precisión	0.5000
Sensibilidad	1.0000
Especificidad	0.0000
F1-score	0.6667
AUC	0.4062
Tiempo (s)	377.01
CO_2 (kg)	0.0101

Cuadro 4.3: Resultados del modelo CNN+GRU con técnicas de aumentación de datos.

Los resultados de esta sección sugieren que la aumentación de datos no aportó beneficios significativos.

4.4. Resultados de *transfer learning*

El objetivo de esta sección era comprobar como el uso de modelos preentrenados en grandes bases de datos de imágenes aumentaría significativamente la capacidad de generalización de los modelos.

La Tabla 4.4 presenta los resultados obtenidos por los distintos modelos CNN preentrenados evaluados en este estudio. MobileNetV2 alcanzó el mejor rendimiento general, con una sensibilidad del 81.25 % y un F1-score de 0.8387, superando al resto de arquitecturas tanto en capacidad de detección de casos positivos como en equilibrio entre precisión y exhaustividad. VGG-16 también mostró un comportamiento competitivo, especialmente en sensibilidad (68.75 %). Por el contrario, ResNet50 fue el modelo con peor rendimiento, con una sensibilidad nula y un F1-score de 0. Las diferencias en el tiempo de entrenamiento entre modelos fueron notables, siendo VGG-16 el más eficiente en términos de velocidad (211.64 segundos), frente al mayor tiempo requerido por MobileNetV2 (1462.86 segundos).

Además, en cuanto a las emisiones de CO_2 asociadas al proceso de entrenamiento de cada modelo, aunque MobileNetV2 fue el modelo más preciso, también fue el que generó un mayor impacto ambiental (0.0302 kg de CO_2). En contraste, VGG-16 presentó una buena relación entre rendimiento y eficiencia energética, con una emisión estimada de solo 0.0073 kg de CO_2,

lo que puede ser relevante en contextos donde la sostenibilidad computacional sea un factor a considerar.

Métrica	**VGG-19**	**VGG-16**	**ResNet50**	**MobileNetV2**
Exactitud	0.5625	0.6250	0.4375	0.8438
Precisión	0.6667	0.6111	0.0000	0.8667
Sensibilidad	0.2500	0.6875	0.0000	0.8125
Especificidad	0.8750	0.5625	0.8750	0.8750
F1-score	0.3636	0.6471	0.0000	0.8387
AUC	0.6719	0.7266	0.3945	0.9023
Tiempo (s)	324.75	211.64	1287.01	1462.86
CO_2 (kg)	0.0119	0.0073	0.0252	0.0302

Cuadro 4.4: Resultados obtenidos con los modelos CNN preentrenados: métricas de evaluación, tiempo de entrenamiento y emisiones estimadas de CO_2.

A pesar de los resultados dispares entre modelos, la implementación de modelos preentrenados obtuvo, en general, resultados significativamente superiores a los modelos base anteriores, con y sin aumentación de datos.

4.5. Resultados de modificaciones estructurales

Dadas las restricciones computacionales que presentaba el proyecto, se implementaron diferentes modificaciones estructurales al modelo preentrenado con VGG-16, ya que ofrecía el mejor balance entre precisión y coste computacional.

La Tabla 4.5 resume los resultados obtenidos tras aplicar las distintas modificaciones estructurales a este modelo. Estas incluyen la técnica de *slice stacking*, la incorporación de un mecanismo de atención y el descongelado parcial de capas convolucionales. El modelo con descongelado de capas mostró el mejor equilibrio general, alcanzando una sensibilidad del 93.75 % y un F1-score de 0.7317. En contraste, el modelo con *slice stacking* logró una sensibilidad perfecta (100 %), pero con una especificidad extremadamente baja (6.25 %), lo que indica un sesgo hacia la clase positiva. En cuanto al tiempo de entrenamiento, todas las variantes se mantuvieron en rangos similares y relativamente bajos, con emisiones de CO_2 también contenidas.

Métrica	Slice stacking	Mecanismo de atención	Descongelado de capas
Exactitud	0.5312	0.6562	0.6562
Precisión	0.5161	0.7273	0.6000
Sensibilidad	1.0000	0.5000	0.9375
Especificidad	0.0625	0.8125	0.3750
F1-score	0.6809	0.5926	0.7317
AUC	0.4531	0.6562	0.6328
Tiempo (s)	186.02	341.95	346.19
CO_2 (kg)	0.0060	0.0129	0.0127

Cuadro 4.5: Resultados de modelos CNN+GRU con modificaciones estructurales: métricas de evaluación, tiempo de entrenamiento y emisiones estimadas de CO_2.

Los resultados obtenidos no mostraron una mejoría significativa frente al modelo preentrenado con VGG-16 sin modificaciones estructurales. Por lo tanto, dado que ninguna de las técnicas logró resultados similares a los obtenidos por el modelo preentrenado con MobileNetV2, se escogió este último para las siguientes secciones.

4.6. Resultados de optimización de hiperparámetros

Con el objetivo de optimizar el rendimiento del modelo preentrenado con MobileNetV2, se llevó a cabo una búsqueda de hiperparámetros mediante la librería *Optuna*. Los hiperparámetros explorados incluyeron 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 del *learning rate*. El mejor conjunto de hiperparámetros obtenido fue el siguiente: *l2* = 4.94e-05, *gru units* = 64, *dense units* = 64, *dropout rate* = 0.289 y *learning rate* = 0.00099.

El modelo entrenado con esta configuración logró una sensibilidad del 81.25 %, con un rendimiento equilibrado en el resto de métricas (precisión, especificidad y F1-score = 0.8125), como se muestra en la Tabla 4.6.

Métrica	Modelo optimizado (Optuna)
Exactitud	0.8125
Precisión	0.8125
Sensibilidad	0.8125
Especificidad	0.8125
F1-score	0.8125
AUC	0.8906
Tiempo (s)	1256.22
CO_2 (kg)	0.0144

Cuadro 4.6: Resultados del modelo CNN+GRU optimizado mediante búsqueda de hiperparámetros.

Al comparar estos resultados con los del modelo MobileNetV2 base de la sección anterior, se observa que ambos alcanzaron la misma sensibilidad (81.25 %). Sin embargo, de forma contraintuitiva, el modelo optimizado mostró un rendimiento menor en el resto de las métricas.

4.7. Resultados de funciones de pérdida robustas

El objetivo de esta sección era examinar si la implementación de diferentes funciones de pérdida produciría una mejora en el modelo MobileNetV2, ya que fue el que mejores resultados obtuvo.

La Tabla 4.7 muestra los resultados obtenidos al aplicar las distintas funciones de pérdida robustas con el objetivo de mitigar el impacto de posibles etiquetas erróneas en el conjunto de datos. Las funciones evaluadas fueron Focal Loss, Label Smoothing e iMAE. De entre ellas, el mejor resultado general fue obtenido por el modelo entrenado con Label Smoothing, que alcanzó una sensibilidad del 75 % y un F1-score de 0.8276. Por su parte, el modelo con Focal Loss mostró una precisión perfecta (1.0000) pero una sensibilidad algo más reducida (56.25 %), mientras que la variante con iMAE ofreció resultados globalmente inferiores. En términos de eficiencia computacional, las emisiones de CO_2 y los tiempos de entrenamiento fueron ligeramente superiores a los de MobileNetV2, especialmente en el caso de Focal Loss.

Métrica	**Focal Loss**	**Label Smoothing**	**iMAE**
Exactitud	0.7812	0.8438	0.5312
Precisión	1.0000	0.9231	0.5294
Sensibilidad	0.5625	0.7500	0.5625
Especificidad	1.0000	0.9375	0.5000
F1-score	0.7200	0.8276	0.5455
AUC	0.9141	0.8789	0.6289
Tiempo (s)	1510.75	1445.36	1281.90
CO_2 (kg)	0.0321	0.0306	0.0273

Cuadro 4.7: Resultados de modelos con funciones de pérdida robustas: métricas de evaluación, tiempo de entrenamiento y emisiones estimadas de CO_2.

Los resultados obtenidos no muestran una mejoría significativa frente al modelo MobileNetV2 base. A pesar de que algunas métricas mejoraban, la sensibilidad, que era la métrica principal en este proyecto, disminuyó en los tres casos.

4.8. Resultados de *Grad-Cam*

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).

De entre las técnicas implementadas, el *transfer learning* fue la que obtuvo los resultados más positivos. En cambio, el *data augmentation*, la implementación de funciones de pérdida robustas, la optimización de hiperparámetros y la aplicación de técnicas de interpretabilidad no dieron los resultados esperados. La causa de estos resultados y sus implicaciones se expondrán en el siguiente capítulo.

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.

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.

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.

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.

Chavez, C., Schurger-Foy, A., Oh, J., Oh, J. K., Linstead, E., Yun, K. & Maoz, U. (2025), 'Toward the trusted medical imaging ai: An explainable machine learning model for schizophrenia brain mris', *SSRN* .

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.

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.

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.

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.

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.

Qian, J., Li, H., Wang, J. & He, L. (2023), 'Recent advances in explainable artificial intelligence for magnetic resonance imaging', *Diagnostics* **13**(9), 1571.

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.

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.

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

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, 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.

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.

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.