Universidad Internacional de La Rioja (UNIR)

ESIT

Máster Universitario en Inteligencia Artificial

Medicina personalizada: comparativa de técnicas para el diagnóstico automático del Alzheimer

Trabajo Fin de Máster

Presentado por: Darias Plasencia, Óscar

Directora: Mancera Valetts, Laura

Ciudad: Santa Cruz de Tenerife
Fecha: miércoles, 24 de julio de 2019

AGRADECIMIENTOS

En primer lugar, agradecer a todo el personal de UNIR involucrado, principalmente mi tutor personal y los profesores de las diferentes asignaturas. Ha sido una gran experiencia que me ha comunicado verdadera pasión por el campo de la Inteligencia Artificial, al que espero poder dedicarme profesionalmente en el futuro. Gracias también a la directora de este TFM, que me ha ayudado constantemente en la redacción de este documento y otras dudas. Finalmente, agradecer a toda mi familia su apoyo incondicional, algunos de ellos no sólo moral sino también en cuestiones sobre medicina. Ha sido una experiencia dura y solitaria en muchos momentos, pero el trabajo ha salido adelante con el apoyo de todas estas personas.

La recolección y cesión de los datos para este proyecto fue realizada por el *Alzheimer's Disease Neuroimaging Initiative* (ADNI) (*National Institutes of Health Grant* U01 AG024904) y DOD ADNI (*Department of Defense award number* W81XWH-12-2-0012). El ADNI es financiado por el Instituto Nacional del Envejecimiento, el Instituto Nacional de Imagen Biomédica y Bioingeniería, y a través de generosas contribuciones de los siguientes organismos: *AbbVie*, Asociación de Alzheimer; *Alzheimer's Drug Discovery Foundation*; *Araclon Biotech*; *BioClinica, Inc*; *Biogen*; *Bristol-Myers Squibb Company*; *CereSpir, Inc*; *Cogstate*; *Eisai Inc*; *Elan Pharmaceuticals, Inc*; *Eli Lilly and Company*; *EuroImmun*; *F. Hoffmann-La Roche Ltd*. y su empresa afiliada *Genentech, Inc*. y *Fujirebio*; *GE Healthcare*; *IXICO Ltd*.; *Janssen Alzheimer Immunotherapy Research & Development, LLC*; *Johnson & Johnson Pharmaceutical Research & Development LLC*; *Lumosity*; *Lundbeck*; *Merck & Co, Meso Scale Diagnostics, LLC*; *NeuroRx Research*; *Neurotrack Technologies*; *Novartis Pharmaceuticals Corporation*; *Pfizer Inc*; *Piramal Imaging*; *Servier*; *Takeda Pharmaceutical Company*; y *Transition Therapeutics*. Los Institutos Canadienses de Investigación en Salud están proporcionando fondos para apoyar a los centros clínicos del ADNI en Canadá. Las contribuciones del sector privado son facilitadas por la Fundación para los Institutos Nacionales de Salud (www.fnih.org). La organización beneficiaria es el Instituto de Investigación y Educación del Norte de California, y el estudio está coordinado por el Instituto de Investigación Terapéutica de Alzheimer de la Universidad del Sur de California. Los datos de ADNI son diseminados por el Laboratorio de Neuroimagen de la Universidad del Sur de California.

Todo el código implementado para el desarrollo de este trabajo, así como todos los informes, se encuentran en el siguiente repositorio: https://github.com/oscardp96/TFM-Alzheimer-Diagnosis

Resumen

En los últimos años, se ha dado un creciente interés por automatizar el diagnóstico por imagen de diferentes enfermedades con Inteligencia Artificial, entre ellas el Alzheimer. Esto permite reducir el error humano y hacer el diagnóstico mucho más preciso, gracias a las capacidades del aprendizaje automático y aprendizaje profundo para encontrar patrones ocultos en las imágenes médicas. En este trabajo, se estudia qué modelos son los más indicados para esta tarea, y se implementan en un escenario experimental por medio de herramientas sencillas, para comprobar hasta qué punto es factible su desarrollo por parte de profesionales sin conocimientos profundos de medicina. También se busca emitir las recomendaciones oportunas. Los resultados muestran que las redes neuronales convolucionales permiten construir modelos bastante potentes, pero su despliegue en entornos reales es poco factible hoy en día.

Palabras Clave: Alzheimer, aprendizaje profundo, diagnóstico, análisis de imagen médica

Abstract

In recent years, there has been a growing interest in automating the diagnosis by image of different diseases with Artificial Intelligence, including Alzheimer. This trend is trying to reduce human error and make the diagnosis much more precise, thanks to the capabilities of machine learning and deep learning to find hidden patterns in medical images. In this work, several state-of-the-art models are studied to find the most suitable ones for this task, and they are implemented in an experimental scenario by means of simple tools, in order to verify to what extent their development is feasible by professionals without deep knowledge of medicine. It also seeks to issue the appropriate recommendations for these professionals. The results show that convolutional neural networks allow the construction of quite powerful models, but their deployment in real environments is hardly feasible today.

Keywords: Alzheimer, Deep Learning, diagnosis, medical image analysis

ÍNDICE

Contenidos

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1 Introducción

Esta primera sección inicia con una visión general de la Inteligencia Artificial (IA) aplicada a la medicina, para luego profundizar en la temática del diagnóstico del Alzheimer usando precisamente IA. Adicionalmente, se presenta el alcance de la investigación, así como las limitaciones que se deben tener en cuenta, no sólo a la hora de desarrollar el trabajo (para el autor) sino también a la hora de leer este documento (para el lector). Por último, se resume la estructura que sigue el resto de este documento.

1.1 Motivación

La Inteligencia Artificial, con los años, se espera que forme parte fundamental de nuestra vida cotidiana (Russell y Norvig, 2016). Especialmente en las últimas dos décadas, se han venido materializando avances teóricos y prácticos de gran relevancia en una amplia variedad de campos, como aplicaciones médicas, turismo, educación, entretenimiento, entre otros (Dilsizian y Siegel, 2014; Gretzel, Sigala, Xiang, y Koo, 2015; Millington y Funge, 2009; Mancera, Baldiris, y Fabregat, 2008). Específicamente, en el campo de la medicina, que es el enfoque general de esta investigación, la IA ha dejado ver su potencial en diversos aspectos como: la asistencia en la formulación de un diagnóstico, la detección de lesiones, o la segmentación de células y órganos (Akkus, Galimzianova, Hoogi, Rubin, y Erickson, 2017; Ciompi et al., 2015; Lo et al., 1995).

La medicina tiene particularidades muy distintivas con respecto a cualquier otro ámbito donde se ha aplicado la IA hasta el momento, aunque al final, los grandes beneficios que se pueden obtener son bastante similares. Estos, a grandes rasgos, son dos: el primero es la automatización de procesos tradicionalmente realizados por los seres humanos (en este caso, el personal clínico), la cual agiliza el trabajo que, originalmente, era realizado por estos en su totalidad. Ejemplos podrían ser la organización automática de la información para su fácil acceso, o la toma de decisiones en base a unas pruebas previamente realizadas. En este ámbito, se destaca el papel de los sistemas cognitivos, los cuales, precisamente, manejan enormes cantidades de información y reglas de inferencia para asistir a los humanos en la toma de decisiones con ayuda de modelos de aprendizaje automático en su entorno laboral. Todo esto con un sistema de entrada / salida amigable para el usuario. Y el segundo rasgo, es la obtención de nuevos métodos automatizados que permiten identificar patrones o realizar tareas que el ser humano no es capaz de ejecutar por sí mismo, o sí es capaz, pero con un amplio margen de error. Este podría ser el caso del análisis de imágenes médicas para tratar de obtener un diagnóstico.

Un lector con cierta experiencia en el campo del aprendizaje automático sabrá que estas mismas ventajas se aplican a la gran mayoría de disciplinas donde la Inteligencia Artificial tiene cabida. De hecho, casi podrían establecerse como las ventajas que la IA proporciona de manera general. Sin embargo, el campo de la medicina plantea una serie de retos y complicaciones adicionales, no sólo técnicas sino también éticas. Hacen que las aplicaciones de la IA sobre esta disciplina se diferencien de otras aplicaciones típicas industriales, donde las herramientas y algoritmos se pueden aplicar de forma más directa. Todos estos retos se abordan en el capítulo 2, dedicada a la revisión del estado del arte.

La Inteligencia Artificial en medicina da una motivación extra muy clara, y es la posibilidad de mejorar o incluso salvar vidas. La perspectiva de trasladar los enormes avances que se han obtenido en la última década, especialmente en el campo del aprendizaje profundo o *Deep Learning*, al campo de la medicina, podría dar lugar a mejoras significativas en el diagnóstico o tratamiento de enfermedades, así como en muchas otras aplicaciones. No obstante, estas técnicas sólo han empezado a tomar importancia en la investigación médica en los últimos cinco años. Aún queda mucho por avanzar, pero sin duda es un campo que cada vez tiene mayor actividad en el que, por suerte, la mayor parte de las investigaciones son de acceso abierto, lo cual otorga muchas esperanzas de progresar significativamente en los próximos años.

Dentro del paradigma de la aplicación de la inteligencia artificial a la medicina, este trabajo se centra en el diagnóstico automatizado del Alzheimer. Esta enfermedad es una de las enfermedades neurodegenerativas más notables en la actualidad y la forma más común de demencia. El aumento de la esperanza de vida en las últimas décadas ha provocado que las muertes por esta enfermedad hayan aumentado un 66% (Thies y Bleiler, 2012), mientras prácticamente la totalidad del resto de causas de muerte han descendido; de hecho, “para 2050, se espera que se dé un caso nuevo de Alzheimer cada 33 segundos” (Thies y Bleiler, 2012, p.1). Se trata de una enfermedad fatal producida por la muerte o mal funcionamiento de las neuronas, llegando al punto en que el paciente es incapaz de llevar a cabo funciones muy básicas, como podrían ser caminar o incluso tragar (Thies y Bleiler, 2012). El problema se agrava debido a que las causas específicas que dan lugar a esta enfermedad son mayormente desconocidas (Reitz, Brayne, y Mayeux, 2011; Thies y Bleiler, 2012). Esto resulta en dificultades para prevenir y tratar la enfermedad.

En este contexto, se han venido realizando esfuerzos apoyados en el aprendizaje automático para identificar si un sujeto puede padecer de Alzheimer o no, usando entre otro tipo de información, datos de Imagen por Resonancia Magnética (IRM). En el mundo del aprendizaje automático, identificar si un sujeto puede padecer de Alzheimer o no, es un

problema de clasificación. Normalmente, suelen hacerse tres distinciones: paciente sano (HC, *Healthy Control*), paciente con deterioro cognitivo leve (MCI, *Mild Cognitive Impairment*) o paciente enfermo (AD, *Alzheimer Disease*).

El auge del *Deep Learning* en los últimos años ha dado lugar a un enorme abanico de investigaciones, en las que se proponen diversas técnicas para tratar de hacer frente a los retos específicos del problema abordado. Además, antes del auge de las redes neuronales convolucionales, otras investigaciones han abordado este mismo problema con otras técnicas, algunas con cierta influencia del *Deep Learning*, y otras con un enfoque más estadístico (Greenspan, van Ginneken, y Summers, 2016; Ker, Wang, Rao, y Lim, 2018; Litjens et al., 2017).

La variedad de opciones es muy grande, y cada una tiene sus propias ventajas e inconvenientes. Además, se sabe desde hace ya más de una década que estos modelos pueden diagnosticar el Alzheimer por medio de imágenes médicas mejor de lo que es capaz de hacerlo un radiólogo experimentado, o al menos de forma comparable (Klöppel et al., 2008). Es decir, que la mayoría de los modelos propuestos en todas estas investigaciones tienen potencial para ser desplegados en entornos clínicos reales. Por tanto, es importante preguntarse, ¿cuál de las opciones ofrece mejor rendimiento, no sólo a la hora de desplegar sistemas en entornos reales sino también a la hora de profundizar en la investigación? Esta, precisamente, es una de las preguntas de investigación que orienta este TFM.

También es importante tener en cuenta que la construcción de estos modelos requiere, en ciertos casos, de conocimientos bastante profundos de medicina. Por ejemplo, para preprocesar las imágenes médicas de la mejor forma posible para aprovechar sus características, son necesarios conocimientos exhaustivos de qué representa una imagen médica T1 ponderada o TEP (tomografía por emisión de positrones). De la misma manera, los algoritmos utilizados antes del auge del aprendizaje profundo requerían de la especificación manual de las características importantes de las imágenes, lo que implica conocimiento de los biomarcadores que sirven para diagnosticar el Alzheimer.

Como es de esperar, la mayoría de los científicos de datos o profesionales dedicados a la creación de modelos de inteligencia artificial no tienen estos conocimientos. Si bien es cierto que un mínimo conocimiento del dominio es siempre necesario para un científico de datos a la hora de desenvolverse, las particularidades de la medicina pueden dificultar mucho esta tarea. Por tanto, este TFM también hace una implementación de los algoritmos para comprobar hasta qué punto la producción de este tipo de modelos es factible para un profesional sin conocimiento profundo sobre el Alzheimer, teniendo simplemente las

nociones básicas necesarias para estudiar el estado del arte y reproducir las principales alternativas.

Cabe destacar que, de las dos principales ventajas de la aplicación de la IA en medicina expuestas en el apartado anterior, el diagnóstico del Alzheimer se beneficia de ambas al mismo tiempo. En primer lugar, sirve para automatizar un proceso que, hoy en día, es realizado por el personal clínico. Pero, al mismo tiempo, el hecho de que las causas específicas que dan lugar a la enfermedad son mayormente desconocidas, hace que se obtenga un mayor beneficio de la identificación de patrones de la que son capaces los modelos de aprendizaje automático. De hecho, los modelos de aprendizaje profundo son incluso capaces de encontrar, por sí mismos, características relevantes que sirvan para el diagnóstico.

1.2 Objetivos

El objetivo general de este trabajo de fin de máster es generar recomendaciones sobre los modelos de aprendizaje automático más indicados para apoyar la toma de decisiones de un sistema de diagnóstico del Alzheimer. Esto implica una revisión bibliográfica y su implementación a partir de una base de datos de neuroimagen.

Los objetivos específicos serían los siguientes:

1. Identificar los modelos que mejores resultados están proporcionando en el diagnóstico del Alzheimer, por medio de una revisión sistemática de la literatura de este siglo.
2. Implementar los modelos seleccionados en el objetivo específico 1, haciendo uso de la base de datos de neuroimagen del ADNI (Jack et al., 2008), con el fin de comprobar cómo de factible es su implementación para profesionales sin conocimientos profundos de medicina.
3. Comparar los resultados obtenidos por los diferentes modelos en el objetivo específico 2, analizarlos y emitir las recomendaciones oportunas.

1.3 Metodología de trabajo

El proceso de investigación en esta tesis se dividió en dos fases principales: la fase exploratoria y la fase descriptiva.

La fase exploratoria se llevó a cabo para identificar aquellos modelos que más se usan en el diagnóstico automatizado del Alzheimer. En particular se realizaron las siguientes actividades en esta fase:

- Elaboración de una revisión sistemática de literatura que permitió una mejor comprensión de la aplicación del aprendizaje automático en el diagnóstico del Alzheimer.
- Identificación de los modelos de aprendizaje automático más ampliamente utilizados en la literatura, así como las principales técnicas de preprocesamiento de neuroimagen para esos modelos. Como resultado, esta actividad proporcionó información sobre los modelos que serían más indicados a la hora de desplegar un sistema de diagnóstico del Alzheimer en un entorno real, y sobre los principales métodos de preprocesamiento de la información necesarios para ponerlos en práctica.
- Selección de dos modelos en base a toda la información disponible hasta este punto.
- Estudio de herramientas que permitan el prototipado de los modelos seleccionados. Se trata de librerías de programación de modelos de aprendizaje automático.

La fase descriptiva se diseñó con el propósito de describir el funcionamiento de los dos modelos seleccionados en la fase anterior y comparar de manera práctica su funcionamiento. Para ello se realizaron las siguientes actividades:

- Preprocesamiento de un conjunto de imágenes de resonancia magnética (IRM), con el objetivo de preparar la información para entrenar modelos de aprendizaje automático. Se ponen en práctica las técnicas de preprocesamiento de la información analizadas durante la fase exploratoria.
- Prototipado de los modelos seleccionados durante la fase exploratoria, haciendo uso de las imágenes preprocesadas y comparando los resultados obtenidos.
- Valoración final teniendo en cuenta los resultados obtenidos y la información extraída durante la fase exploratoria.
- Realizar recomendaciones pensadas para profesionales de Inteligencia Artificial sin conocimientos profundos de medicina. Partiendo de las conclusiones de la actividad anterior, se busca proporcionar una base sobre la que partir a la hora de construir modelos de diagnóstico.

1.4 Alcance y limitaciones

Este trabajo no busca hacer modificaciones a las soluciones propuestas en el estado del arte. En su lugar, se pretende estudiar cuáles son los modelos propuestos (por ejemplo, redes neuronales convolucionales) que más importancia están teniendo, e implementarlos de un modo más directo, haciendo uso de las librerías disponibles y comúnmente utilizadas por los profesionales dedicados a la implementación de modelos de aprendizaje automático. De esta manera, se determina si estas herramientas son suficientes para crear buenos modelos, sin necesidad de que el profesional tenga conocimientos profundos de medicina.

Tampoco se pretende adoptar un enfoque puramente clínico, en el sentido de que no se estudiarán los biomarcadores específicos para el diagnóstico del Alzheimer. Esto, al contrario de lo que podría parecer, no es un problema mayor que impida implementar una solución. Como se explica en el apartado 2.4, el aprendizaje profundo o *Deep Learning* tiene la ventaja de que los propios modelos desarrollados son capaces de identificar los biomarcadores por sí mismos. Esto contrasta con las soluciones más comunes de finales de la década de los 2000 y principios de la de 2010, donde en muchos casos se hace necesaria la especificación manual de características relevantes. Estas características serían precisamente los biomarcadores, en los cuales se profundiza ligeramente en el apartado 2.2. No obstante, y pese a ello, estos modelos obtienen peores resultados, así que no serán parte de la experimentación.

Este último punto incluye la estructura de las imágenes médicas. La variedad de tipos es muy grande (PET, T1, T2…), teniendo cada una sus propias particularidades. Hay que tener en cuenta que no se trata de imágenes corrientes en dos dimensiones, sino imágenes en tres dimensiones, con una profundidad que puede superar muy fácilmente las 70 superficies. En general, cada investigación lleva a cabo su propia forma de preprocesamiento para adaptar estas imágenes a los modelos desarrollados. En este trabajo no se busca aportar ningún tipo de perspectiva novedosa en su estudio o tratamiento.

Al margen de las imágenes médicas, se consideró la posibilidad de combinar la información proporcionada por estas con información de un historial clínico. Esta última es más sencilla de tratar y podría proporcionar características interesantes. Sin embargo, las principales bases de datos de neuroimagen no proporcionan el historial clínico de los pacientes, por lo que se tuvo acceso únicamente a las imágenes y a la edad del paciente.

Más allá de esto, se encuentran las que probablemente son las limitaciones más importantes: el tiempo y la capacidad de procesamiento. No se dispone de equipos excesivamente potentes, ni siquiera de una GPU. La mayor parte del trabajo se ha llevado a

cabo en un *MacBook Pro Retina* de 13 pulgadas de mediados de 2014, con un procesador *Intel Core* i5 a 2.6GHz, memoria de 8GB 1600MHz DDR3 y gráficos integrados *Intel Iris* de 1536MB. Estas especificaciones son un cuello de botella enorme para el desarrollo de redes neuronales que trabajen con imágenes, por lo que el uso de este equipo, al menos de forma directa, no es siquiera factible.

En su lugar, se ha recurrido a los cuadernos de Google Colaboratory para llevar a cabo el diseño y entrenamiento de los modelos. Se trata de un entorno en la nube que permite trabajar con cuadernos de Jupyter (Kluyver et al., 2016). Este entorno ofrece, de forma gratuita, máquinas virtuales que hacen uso de tarjetas gráficas Tesla K80, con 12.72GB de RAM y 350GB de disco duro. También permite interactuar con almacenamiento en Google Drive, donde se han alojado los datos utilizados para entrenar los modelos («Colaboratory: Frequently asked questions», 2018). Al mismo tiempo, se mantienen las capacidades de visualización de datos y exposición de resultados que caracterizan a los cuadernos de Jupyter.

Aunque las especificaciones descritas permiten el entrenamiento de redes neuronales complejas en un tiempo asumible, hay que tener en cuenta que su uso se encuentra ligeramente limitado. El tiempo máximo de conexión a una GPU es de 12 horas, y está pensado exclusivamente para uso interactivo. Es decir, que no es posible realizar extensas búsquedas de hiperparámetros ni entrenamientos de varios cientos de épocas en segundo plano. Por esta razón, los resultados finales analizados en este documento no están soportados de forma consistente por varios entrenamientos diferentes, sino que simplemente dan una idea de los rendimientos que se pueden llegar a obtener.

Por último, es importante dejar claro que los modelos de aprendizaje automático para el diagnóstico médico aún no están listos para su despliegue en entornos reales. Este no solo es el caso de modelos avanzados de diagnóstico del Alzheimer, como el de Ding et al., (2018), sino también de otros modelos exitosos empleados para el diagnóstico de otras enfermedades (Esteva et al., 2017; Gulshan et al., 2016; Vesal, Ravikumar, Davari, Ellmann, y Maier, 2017). Estos modelos no son totalmente fiables hoy en día, y se requiere de más y mejores investigaciones que prueben si es factible llevarlos a producción.

1.5 Estructura de la memoria

Este documento se estructura de la forma descrita a continuación, siguiendo la metodología explicada en la sección 1.3. En la sección 2, se lleva a cabo un proceso sistemático de

revisión de la bibliografía del diagnóstico del Alzheimer basado en Inteligencia Artificial. Engloba la fase exploratoria de esta tesis:

- En el apartado 2.1, se describen los criterios de búsqueda.
- En el 2.2, se lleva a cabo una introducción al análisis de imágenes médicas de forma general, profundizando posteriormente en el papel concreto de la inteligencia artificial (2.2.1).
- En el 2.3, se aborda el tema del diagnóstico del Alzheimer en concreto, destacando sus particularidades.
- En el 2.4, se analiza detenidamente el papel del aprendizaje profundo en la investigación médica, destacando las principales publicaciones que se deben tener en cuenta y describiendo los principales modelos utilizados.
- En el 2.5, se resumen las principales técnicas de preprocesamiento de imágenes médicas.
- En el 2.6, se resumen los principales retos a los que se enfrentan los investigadores.
- La sección se cierra en el apartado 2.7 con las conclusiones obtenidas de la revisión de la bibliografía.

En la sección 3, se describe la fase descriptiva de esta tesis, estructurada en los siguientes apartados:

- En el apartado 3.1, se describe el conjunto de datos utilizado. Esto incluye las fuentes de las que se obtuvieron, en qué condiciones y qué técnicas de preprocesamiento se aplicaron para prepararlos para las siguientes actividades.
- En el apartado 3.2, se describen los diferentes modelos implementados y se resumen brevemente los resultados obtenidos.
- En el apartado 3.3, se lleva a cabo un análisis pormenorizado de los resultados, contextualizando con los obtenidos en otras investigaciones similares.
- En el apartado 3.4, se emiten las recomendaciones oportunas para la implementación de modelos de este tipo, a partir de las conclusiones del análisis anterior.

Finalmente, en la sección 4 se cierra este documento con las conclusiones extraídas del trabajo. También se resumen líneas de trabajo futuras que quedan pendientes a partir de los resultados de esta tesis.

2 Modelos y estado del arte

En esta sección, se lleva a cabo un resumen detallado del estado del arte en el contexto del diagnóstico automatizado del Alzheimer. Inicialmente, se describen el proceso y los criterios seguidos para la revisión de la bibliografía. Seguido, se introduce el concepto del análisis de imágenes médicas, enfatizando en el papel que ha jugado la Inteligencia Artificial en ello en los últimos años. Posteriormente, se expone la información básica relevante sobre el diagnóstico del Alzheimer en concreto, antes de profundizar en los modelos de *Deep Learning* más relevantes en este campo. Por último, se analizan los principales retos a los que se enfrentan los investigadores hoy en día en este campo de la medicina.

Los estudios referenciados en esta sección serán fundamentales para los siguientes capítulos, pues de ellos parte el trabajo realizado. Este era, además, el primero de los objetivos de este TFM, y comprende la fase exploratoria de la metodología de investigación. Es importante tener muy en cuenta que, aunque la mayoría de los puntos que se presentan son comunes a la totalidad de las aplicaciones de la Inteligencia Artificial sobre el análisis de imágenes médicas, los ejemplos e investigaciones citadas tratan el diagnóstico del Alzheimer.

2.1 Proceso de revisión de bibliografía

El proceso de revisión de bibliografía se ha llevado a cabo siguiendo un proceso sistemático, para asegurar la lectura de las publicaciones más relevantes del estado del arte. El buscador utilizado ha sido Google *Scholar*, ordenando cada búsqueda realizada siempre por relevancia y prestando atención al índice de impacto.

Centrados directamente en el campo de la medicina, se estudiaron 4 artículos de revista. Estos proporcionaron información sobre la propia enfermedad del Alzheimer y sobre la importancia de los biomarcadores. Más allá de estos, la revisión se centró completamente en aplicaciones de aprendizaje automático.

Al ser un campo relativamente pequeño, es decir, que el número de investigaciones publicadas no es excesivamente grande, los artículos importantes se citan entre sí, simplificando la revisión de la bibliografía. Por ejemplo, la publicación de Sarraf, Tofighi, y otros (2016) incluye en el apartado 3 un resumen del estado del arte hasta el año 2016, así como una tabla con las investigaciones previas más importantes y sus resultados.

La mayoría de estas investigaciones, además, requerían de conocimientos sobre ciertos modelos de *machine learning* que eran desconocidos para el autor. Por ejemplo, los autoencoders y las RBM (*Restricted Boltzmann Machines*). Esto obligó a realizar la búsqueda de los artículos originales que presentaban estos modelos, para poder entenderlos y, posteriormente, hacer uso de algunos de ellos. En total, se identificaron 18 trabajos en este grupo, 6 actas de conferencia y 12 artículos de revista.

También es importante destacar el papel fundamental que han tenido las investigaciones generales dedicadas totalmente a recopilar el estado del arte del análisis de imágenes médicas con *Deep Learning* (Greenspan et al., 2016; Ker et al., 2018; Lakhani, Gray, Pett, Nagy, y Shih, 2018; Litjens et al., 2017; Razzak, Naz, y Zaib, 2018; Shen, Wu, y Suk, 2017). Se obtuvieron como resultado de búsquedas por medio de palabras clave como *Deep Learning diagnosis* o *Deep Learning medical image analysis*. Fueron muy útiles no sólo para introducirse al análisis de imágenes médicas en general, sino que además hacen referencia a las técnicas más importantes que hay que conocer, así como a las investigaciones más relevantes que se deben tener en cuenta.

Con las publicaciones citadas en estos artículos, junto con búsquedas en Google *Scholar* por medio de palabras clave como *Deep Learning Alzheimer*, se obtuvieron finalmente 21 trabajos directamente relacionados con el diagnóstico del Alzheimer, 17 de ellos artículos de revista, 3 actas de congreso (*conference proceeding*) y 1 sección de libro (*book section*). Todos ellos posteriores al año 2004, y siendo el más reciente de 2018.

Los trabajos anteriores se analizaron de acuerdo con una serie de categorías, reflejadas en las diferentes secciones de este capítulo. Se buscaba llevar a cabo un análisis estructurado, con conceptos de más generales a más concretos. De esta manera, se comienza estudiando el análisis de imágenes médicas con aprendizaje automático en general, para luego entrar en el diagnóstico del Alzheimer en concreto. Tras el análisis de los diferentes modelos utilizados comúnmente, se profundiza en aquellos que hacen uso de aprendizaje profundo.

Cabe destacar que, de los trabajos más recientes (2016-2018), se extrajeron 3 artículos de revista centrados en aplicaciones de *Deep* Learning para enfermedades diferentes al Alzheimer. Se buscaban puntos en común y la posibilidad de explorar otras ideas.

A continuación, se presenta el análisis realizado por cada una de las categorías mencionadas en el párrafo anterior.

2.2 Análisis de imágenes médicas

Las imágenes médicas representan un aspecto fundamental para el diagnóstico y tratamiento de múltiples enfermedades. Se trata de una fuente muy útil de donde pueden extraerse biomarcadores de diversas patologías (O'Connor et al., 2017). Los biomarcadores son evidencia clínica objetiva medible, y podría decirse que representan la "evidencia clínica más objetiva y cuantificable que un laboratorio científico moderno nos permite medir de forma reproducible" (Strimbu y Tavel, 2010, p. 2). Esto implica que también son indicadores de otros procesos, y no necesariamente de enfermedades, pero interesa centrarse exclusivamente en estos para el trabajo.

Este tipo de imágenes son parte importante de los Expedientes Clínicos Electrónicos (ECE) o, en inglés, *Electronic Health Records* (EHR). Son examinadas normalmente por radiólogos, profesionales dedicados, y las hay de múltiples tipos. En neuroimagen, es decir, lo relativo a las imágenes médicas del cerebro, existe una gran variedad de escaneos diferentes que se pueden realizar. Por ejemplo, en la base de datos OASIS-3 de OASIS *Brains* (OASIS Brains, 2018) se ofrecen 13 tipos de escaneos diferentes para descargar: BOLD, PET, FLAIR, T1w, T2star o T2w, entre otros.

Con el tiempo, ha aparecido un interés creciente en procesar las imágenes médicas de forma automática, lo cual trae consigo una gran cantidad de ventajas, entre las que se encuentran las mencionadas en la sección 1.1 (motivaciones). De hecho, se espera que la inteligencia artificial llegue a jugar un papel muy importante en el análisis de los ECE en general, no sólo de las imágenes (Ker et al., 2018).

2.2.1 El papel de la inteligencia artificial

Dentro de la historia de la Inteligencia Artificial, los sistemas expertos podrían considerarse el primer gran despliegue de tecnologías de este tipo en múltiples industrias. Se trata de la década de los 70 y del paradigma de la Inteligencia Artificial simbólica. En esta época también llegaron los primeros de estos sistemas a la medicina (Ker et al., 2018), aunque, como era de esperar, estaban extremadamente limitados.

Más recientemente, el uso de algoritmos y sistemas de IA ha aumentado considerablemente. Esta fuerte irrupción en el procesamiento de los ECE a diferentes niveles se podría atribuir, principalmente, a tres factores:

- El crecimiento del número y tamaño de estos registros, así como la calidad de su información. Poco a poco, este entorno se aproxima al *Big Data*. Tal y como ha

ocurrido en la mayoría de las áreas donde este paradigma ha irrumpido, la Inteligencia Artificial se convierte en una herramienta idónea para gestionar la información de forma eficiente (Razzak et al., 2018). Por el contrario, a los seres humanos se nos va haciendo cada vez más complicado abarcar tanta información.

- El error humano. Un radiólogo, al igual que cualquier profesional, está limitado por aspectos como la velocidad o la experiencia, y puede cometer errores o sesgar sus decisiones (Ker et al., 2018; Klöppel et al., 2008). De hecho, según Matsuda (2007), incluso los más experimentados fallan entre un 10% y un 15% de las veces.

- El auge de la propia Inteligencia Artificial y, más concretamente, del *Deep Learning*. Si este tipo de sistemas no hubiesen demostrado un rendimiento tan destacado en las últimas décadas, sobre todo tras el trabajo de LeCun, Bottou, Bengio, Haffner, y otros (1998), primero, y de Krizhevsky, Sutskever, y Hinton (2012), más recientemente, llevar estas técnicas a la medicina no se hubiese visto como algo factible. Al menos, no para realizar tareas tan complejas.

Este último punto también conlleva muchos otros factores, como los avances en unidades de procesamiento gráfico, del inglés *Graphics Processing Unit* (GPU). Sin embargo, se asocian con el auge del *Deep Learning* en general, no con el de su uso en aplicaciones médicas.

En los últimos años, la mayor parte de las investigaciones realizadas en diversos ámbitos del análisis de imágenes médicas han apostado por redes neuronales convolucionales (Ker et al., 2018; Litjens et al., 2017). Si bien es cierto que, en el caso concreto del diagnóstico del Alzheimer han aparecido múltiples modelos con buenos resultados basados en máquinas de vector de soporte (SVM), la mayoría de estos utilizan alguna forma de aprendizaje profundo para la extracción de características.

Los usos que se están haciendo de los diferentes modelos también son muy diversos, y tocan diferentes aplicaciones dentro del análisis de imágenes médicas (Greenspan et al., 2016; Ker et al., 2018; Litjens et al., 2017; Shen et al., 2017). A continuación, se van a resumir algunos de ellos, comenzando por la detección de estructuras en imagen (*Computer-aided Detection* - CADe). Se trata de identificar diversos elementos en las imágenes, como órganos o incluso células. También es útil para señalar áreas de interés para un médico, como podrían ser lesiones.

La detección de estructuras no debe confundirse con la segmentación, basada en aislar regiones de la imagen del resto. Pensemos en una imagen corriente de una carretera; la

detección consistiría en simplemente reconocer un coche, mientras que la segmentación requeriría aislarlo del resto de la imagen. En este contexto, la segmentación incluiría una detección previa, y un posterior reconocimiento de los límites. Es útil, por ejemplo, para aislar el cerebro en neuroimagen, y así desechar la información no relevante, como podría ser parte del cráneo.

Por último, dentro de los principales usos, se encuentra el diagnóstico asistido (*Computer-aided Diagnosis* - CADx), también llamado simplemente clasificación. Este trabajo está centrado en este tipo de usos, el cual se basa en llevar a cabo un diagnóstico en base a cierta información. Es decir, a partir de una serie de datos de entrada, el objetivo es clasificarlos en una determinada clase. En ciertos casos, la clasificación es binaria (enfermo o no enfermo), pero en otros puede haber más de dos clases posibles. En el apartado 2.3 se explican cuáles son las clases para este trabajo.

Cabe mencionar que en la bibliografía se pueden encontrar otros usos minoritarios, pero potencialmente útiles. No se entra en detalles sobre todos ellos, pero sí hay uno que merece la pena mencionar: la categoría de aprendizaje profundo de características (Shen et al., 2017). Este se basa en el diseño de sistemas que sean capaces de extraer características útiles de los datos, las cuales suelen ser reutilizables en diferentes escenarios. Esto es una clara ventaja con respecto a las características definidas de forma manual, las cuales no suelen ser reutilizables, y además ligeros cambios en los datos podrían dejarlas inservibles.

Esta última aplicación tiene la particularidad de ser un paso previo útil a las anteriores, como el diagnóstico o la segmentación. En muchos casos, los datos disponibles no son suficientes para entrenar ciertos modelos complejos, por lo que una correcta extracción de características que facilite el entrenamiento se vuelve fundamental (apartado 2.4.1).

En resumen, la tendencia actual es hacer uso de redes neuronales convolucionales (CNN) para realizar aprendizaje supervisado. El objetivo es segmentar regiones, detectar elementos o llevar a cabo un diagnóstico. Para facilitar el entrenamiento de los modelos, en ocasiones también se lleva a cabo una extracción de características previa.

2.3 Diagnóstico del Alzheimer

En el apartado 1.1, se hizo una breve introducción de los principales síntomas que aparecen como resultado de padecer Alzheimer. Estos se asocian principalmente con la memoria, pero resultan en incapacidad para realizar las tareas más básicas (Reitz et al., 2011; Thies y Bleiler, 2012). Evidentemente, cuanto mayores son los síntomas, más sencillo es realizar el

diagnóstico. Sin embargo, existen una serie de pruebas básicas que los médicos pueden realizar para extraer conclusiones sobre si se padece o no la enfermedad, sin necesidad de que aparezcan síntomas avanzados (Thies y Bleiler, 2012).

En 1984 se establecieron los primeros criterios oficiales para llevar a cabo el diagnóstico del Alzheimer; son los llamados criterios NINCDS-ADRDA (Reitz et al., 2011; Thies y Bleiler, 2012). Estos criterios incorporaban biomarcadores que podían asociarse con la enfermedad (Thies y Bleiler, 2012). Sabiendo que los biomarcadores son evidencia clínica medible, tal y como se expuso en el apartado 2.2, es lógico pensar que un sistema de IA debería identificarlos para poder clasificar un paciente como enfermo o sano.

Estos biomarcadores se basan normalmente en medidas de líquido cefalorraquídeo; en indicadores de atrofia en la zona medial del lóbulo temporal en imágenes MRI; o en hipometabolismo en regiones localizadas del cerebro, observable por medio de imágenes TEP (Reitz et al., 2011). Al ser observables por medio de neuroimagen, no es sorprendente que estos biomarcadores se utilicen mucho en el diagnóstico automático, especialmente en los primeros acercamientos de principios de siglo.

Para el diagnóstico CADx del Alzheimer, normalmente se clasifica un paciente en una de tres clases posibles: paciente sano o HC (*healthy control*), paciente con deterioro cognitivo leve o MCI (*mild cognitive impairment*), o paciente enfermo o AD (*Alzheimer disease*). La primera y última de estas clases son triviales. Con respecto a la categoría MCI, se trata de un concepto algo heterogéneo, pero representa un estado que precede al Alzheimer (Matsuda, 2007; Reitz et al., 2011). Se considera que los pacientes MCI tienen un alto riesgo de padecer la enfermedad (Reitz et al., 2011; Thies y Bleiler, 2012). Concretamente, “aproximadamente un 10-15% de los casos de MCI, principalmente pacientes con MCI amnésico, derivan en enfermos de Alzheimer en un año” (Nordberg, Rinne, Kadir, y Lngström, 2010, p. 79).

En la **Tabla *1*** pueden observarse las principales investigaciones que hacían uso directamente de los biomarcadores presentes en las imágenes médicas. La única de estas publicaciones donde se hace uso de aprendizaje automático es en Klöppel et al. (2008). Esta fue la primera investigación relevante que hacía uso de técnicas de este tipo para llevar a cabo el diagnóstico automático del Alzheimer, demostrando que estas técnicas igualaban o superaban a los radiólogos por medio de una comparación directa.

Tabla 1.
Biomarcadores y métodos utilizados para el diagnóstico del Alzheimer en publicaciones hasta 2011.

PUBLICACIÓN	BIOMARCADOR	MÉTODO UTILIZADO
(Imabayashi et al., 2004)	rCBF	3D-SSP
(Matsuda, 2007)	rCBF + Reducción en el metabolismo de la glucosa	Análisis estadístico de imagen
(Klöppel et al., 2008)	Reducción de la materia gris (GM)	SVM
(Yang et al., 2011)	Reducción de la materia gris (GM)	ICA

SVM=Support Vector Machine / Máquina de vector de soporte; rCBF=reducción del flujo sanguíneo cerebral regional; ICA=Independent Component Analysis / Análisis de componentes independientes; 3D-SSP=3-Dimensional Stereotatic Surface Projection / Proyección de superficie estereotáctica tridimensional

A partir del año 2011, comenzaron a utilizarse otros métodos para identificar características que permitieran realizar el diagnóstico, sin especificar los biomarcadores de forma manual. M. Liu, Zhang, Shen, y Initiative (2012) hicieron uso de parches aleatorios extraídos de las imágenes MRI, un método bastante común en el que se entra en detalle en el apartado 2.6.1. Con estos parches, montaron un *ensemble* (combinación) de múltiples modelos débiles basados en clasificación por medio de representaciones dispersas (SRC, *Sparse Representation-based Classification*).

Posteriormente, se comenzaría a hacer uso de aprendizaje profundo para permitir a los modelos identificar por sí mismos las características más relevantes de las imágenes. Este tipo de métodos aparecen en prácticamente todas las investigaciones relevantes posteriores a 2012, comenzando con Gupta, Ayhan y Maida (2013), que utilizaban un autoencoder, y con H. Il Suk y Shen (2013), que utilizaban autoencoders anidados. Eso sí, algunas de estas publicaciones, como esta última, utilizaban las características extraídas para entrenar modelos no profundos, como las SVM.

También hay algunas excepciones recientes que no hacen uso de aprendizaje profundo, sino que llevan a cabo un proceso manual de extracción de características, para luego combinarlas y entrenar una o varias SVM (Mingxia Liu, Zhang, Adeli, y Shen, 2016; Tong et al., 2014; Zu et al., 2016). Otras excepciones hacen uso de métodos radicalmente diferentes, como el análisis discriminante gaussiano (Fang et al., 2017).

En general, estos métodos han sido minoritarios, y la mayor parte de la bibliografía se ha centrado en el aprendizaje profundo.

2.4 *Deep Learning*

En un enorme abanico de tareas, las redes neuronales profundas son el modelo de aprendizaje automático que está logrando mayores avances en los últimos años. Con lo expuesto hasta ahora, se ha dejado ver que esto también se aplica al análisis de imágenes médicas.

Concretamente, al igual que en visión por computador, son las redes neuronales convolucionales las que resultan más prometedoras. Ahora bien, también ha sido fundamental el aprendizaje no supervisado en los últimos años, con modelos como los autoencoders para la extracción de características. En este apartado, se describen con algo más de detalle todos estos modelos. Incluso se explica brevemente su funcionamiento, para lectores no experimentados. Se comienza con las redes neuronales, que son el modelo en el que se basan todos los demás.

Una red neuronal, a grandes rasgos, es un conjunto de unidades o neuronas artificiales que computan una determinada función de los valores que reciben por sus entradas, y envían el valor resultante de la función a través de su salida. Las entradas tienen asignados unos determinados pesos w_i y unos valores extra llamados *biases* b. Para un determinado número de entradas n:

$$y = f(\sum_{i=0}^{n} w_i x_i + b)$$

La f es la función de activación, una función diferenciable que permite introducir cierta no-linearidad en las secuencias tratadas por la red neuronal. Las neuronas se organizan en una determinada topología formada por capas, normalmente conectando todas las neuronas de una capa con todas las neuronas de la siguiente (*fully connected network*). La capa de entrada recibe el *input* (texto, imágenes, audio, etc.), una o más capas ocultas procesan la información, y finalmente una capa de salida emite el resultado.

Estas redes se entrenan de forma supervisada, por medio un algoritmo conocido como *backpropagation*. Este consiste en ir modificando los pesos w y los biases b en sucesivas iteraciones (*batches* y *epochs*), buscando minimizar una función de coste. Se apoya en una técnica conocida como descenso del gradiente (*gradient descent*), y el concepto de las derivadas parciales.

Partiendo de las ideas que hacen funcionar las redes neuronales, se pueden construir varios tipos de modelos. Algunos de los más importantes son los modelos no supervisados, que en el diagnóstico del Alzheimer se utilizan para extracción de características.

2.4.1 Extracción de características

En el apartado 2.2.1, se estableció que una de las razones del auge de la Inteligencia Artificial en medicina era la creciente cantidad de datos disponibles. Sin embargo, como se expone en el apartado 2.6.1, estos siguen sin ser lo bastante numerosos como para alimentar redes neuronales muy profundas. Por tanto, el aprovechamiento de los datos disponibles se vuelve una tarea fundamental para sacar el máximo rendimiento de los modelos y no caer en *overfitting*.

La idea es hacer que los modelos tengan cierta información de antemano acerca de qué características (*features*) tienen que buscar en las imágenes. Esto se consigue por medio de técnicas de aprendizaje profundo, donde se busca que el modelo aprenda cuáles son las características más representativas de las imágenes. Se espera que estas características contengan información clave, que además permita diferenciar unas imágenes de otras. Se hace referencia al aprendizaje no supervisado, pues no se busca realizar ningún tipo de clasificación ni asociar características con una clase en concreto. La propia imagen es todo lo que necesitan estos modelos. Las opciones son muy variadas dentro del campo de la medicina, pero este apartado se centra en las más utilizadas para asistir el diagnóstico del Alzheimer. En la Tabla 2, se presentan las principales publicaciones que hacen uso de este tipo de modelos de aprendizaje no supervisado.

Tabla 2.
Publicaciones que hacen uso de aprendizaje profundo para extracción de características

PUBLICACIÓN	MODELO UTILIZADO
(Gupta et al., 2013)	Autoencoder
(H. Il Suk y Shen, 2013)	Stacked Autoencoders
(H.-I. Suk, Lee, Shen, Initiative, y otros, 2014)	Deep Boltzmann Machines
(H.-I. Suk, Lee, Shen, Initiative, y otros, 2015)	Stacked Autoencoders
(S. Liu et al., 2015)	Stacked Autoencoders
(Payan y Montana, 2015)	Autoencoder
(Li et al., 2015)	Restricted Boltzmann Machines

(Hosseini-Asl, Gimel'farb, y El-Baz, 2016)	Convolutional Autoencoders
(Hosseini-Asl, Keynton, y El-Baz, 2016)	Convolutional Autoencoders

AUTOENCODERS (AE)

En primer lugar, se encuentran los autoencoders. Estos son básicamente redes neuronales en las cuales la entrada es la misma que la salida, por lo que las capas de entrada y salida tendrán el mismo número de neuronas. También suelen tener una única capa oculta. La red primero codifica (*encoder*) la entrada de una determinada manera, y luego decodifica (*decoder*) la representación interna para reconstruir la información original. La calidad del modelo vendrá por tanto determinada por su capacidad para reconstruir la entrada original. El entrenamiento se hace por medio de *backpropagation* y técnicas derivadas del descenso del gradiente, como ocurre con las redes neuronales corrientes.

En la Ilustración 1 se muestra la estructura de un autoencoder básico. Obsérvese la única capa oculta, y las regiones denotadas como *encoder* y *decoder*. La capa de la izquierda sería la capa de entrada, y la de la derecha la de salida.

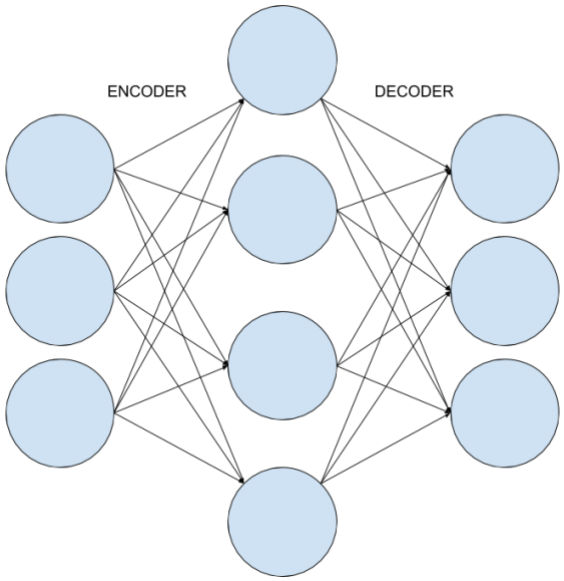

Ilustración 1. Autoencoder básico

Lo interesante de estas redes es obtener los valores que maximizan la activación de cada una de las neuronas de la capa oculta. Estos valores nos permiten obtener una imagen que representa la característica que la neurona está buscando (que podría ser un borde, una línea recta u otro elemento), y que la red ha considerado relevante para reconstruir las imágenes originales. Por cada neurona se obtiene un filtro convolucional que encuentra características relevantes en las imágenes con las que se ha entrenado el autoencoder.

Evidentemente, utilizar valores como infinito o menos infinito da como resultado la máxima activación de la neurona, pero el filtro no tendría sentido. Por ello, se añade la restricción de que el sumatorio de los cuadrados de los valores de activación sea menor o igual que uno. Siendo a_i los N diferentes valores de activación:

$$\sum_{i=0}^{N} a_i^2 \leq 1$$

Además de esta restricción a la hora de obtener las representaciones de las características aprendidas, también hay que añadir otras para evitar que este tipo de modelos aprendan simplemente la función identidad. Esto sería totalmente inútil y no permitiría obtener una codificación de características interesante. Para evitar que esto ocurra, se introduce una nueva restricción: forzar que las neuronas estén inactivas (salida cercana a -1) la mayor parte del tiempo. Esto se logra por medio de un parámetro que modifica la actualización de los *biases* de la red.

El número de unidades de la capa oculta es un hiperparámetro de la red. En general, cuantas más se utilicen, mayor número de filtros se podrán obtener. Sin embargo, al mismo tiempo, aumenta la capacidad de la red y es más sencillo caer en *overfitting*.

Al final, por cada una de las neuronas de la capa oculta, se obtiene un filtro convolucional que, potencialmente, servirá para detectar características interesantes en las imágenes. Por tanto, se espera que sean útiles en las capas convolucionales de una CNN. Múltiples autores utilizan un autoencoder para extraer filtros en 2D (Gupta et al., 2013) o en 3D (Payan y Montana, 2015), para luego utilizarlos en la capa convolucional de una CNN simple. Cabe destacar que, aunque en la bibliografía del diagnóstico del Alzheimer no se han explorado, existen algunas variantes interesantes de los autoencoders. Una de estas son los *denoising autoencoders* o DAEs, que funcionan exactamente igual que los autoencoders básicos, pero con una ligera modificación: se aplica ruido artificial (gaussiano, por ejemplo) a la entrada (Vincent, Larochelle, Lajoie, Bengio, & Manzagol, 2010). De esta forma, el autoencoder tiene que reconstruir la imagen limpia proporcionada como salida, a partir de una imagen ruidosa proporcionada como entrada.

En general, esta técnica sirve para evitar que se aprenda la función identidad, al mismo tiempo que se logra que el autoencoder sea más resistente al ruido, o que incluso sea capaz de reconstruir imágenes ruidosas. Sin embargo, las restricciones descritas anteriormente son más utilizadas en la bibliografía, posiblemente debido a que las IRM no suelen tener problemas de ruido matemático.

AUTOENCODERS ANIDADOS (SAE)

Los autoencoders pueden anidarse para aprender filtros que extraigan características de más alto nivel. La idea es que los filtros de una primera capa servirían para extraer características de bajo nivel, y sucesivas capas proporcionarían filtros para características de más alto nivel. En la Ilustración 2 se muestra la estructura resultante de tres autoencoders anidados.

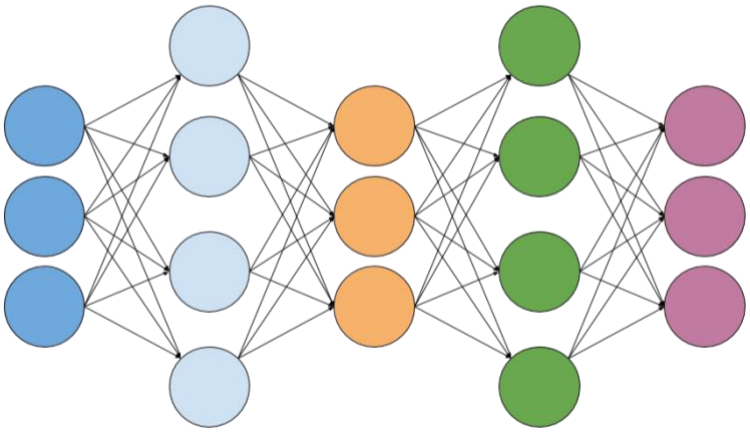

Ilustración 2. Stacked Autoencoder

De esta forma, surgen los autoencoders anidados o *stacked autoencoders* (SAE), los cuales añaden más capas para aprender filtros sucesivos. Es importante no confundir los autoencoders anidados con los autoencoders profundos: los primeros se basan en el encadenamiento de autoencoders individuales, mientras que los segundos consisten en un único autoencoder con múltiples capas ocultas. Esta diferencia se refleja principalmente en el algoritmo de entrenamiento, que en los SAE se conoce como entrenamiento voraz por capas o *greedy layer-wise training* (Hinton, Osindero, y Teh, 2006). Se ha demostrado que esta forma de entrenamiento es útil en autoencoders y otras redes profundas (Bengio, Lamblin, Popovici, y Larochelle, 2007). En la Ilustración 2, se puede observar una serie de capas por colores. La idea es entrenar la primera capa azul clara como se entrena un autoencoder básico, y luego entrenar la capa naranja de la misma forma, pero con la salida de la capa azul clara. El proceso se repetiría para la capa verde, entrenándola con la salida de la capa naranja. De esta forma, se entrena cada capa como un autoencoder individual, a partir de la salida de la capa anterior ya entrenada.

Los filtros obtenidos a partir de un SAE se pueden utilizar de diversas formas. En algunos casos, se utilizan directamente para extraer características, y luego utilizar estas de forma manual para realizar el entrenamiento y clasificación de un modelo. Múltiples autores hacen uso de máquinas de vector de soporte con un kernel basado en estas características para realizar el diagnóstico del Alzheimer (S. Liu et al., 2015; H. Il Suk y Shen, 2013).

Por otro lado, H.-I. Suk et al. (2015) utilizan las sucesivas capas de un SAE para inicializar una red neuronal. Se basan en la idea de que existe información de alto nivel inherente a las características de más bajo nivel, por lo que los SAE son el modelo indicado. Hacen uso de ese entrenamiento no supervisado por capas para obtener unos primeros pesos, y luego añaden una capa final para clasificación y hacen *fine-tuning* de toda la red con los datos supervisados. También hacen uso de una SVM multi-kernel.

SISTEMAS BASADOS EN ENERGÍA

Los *energy-based systems* son otra forma de aprendizaje no supervisado. Concretamente, se han utilizado dos variantes de las máquinas Boltzmann o *Boltzmann Machines* (Ackley, Hinton, y Sejnowski, 1985): las *Restricted Boltzmann Machines* (RBM) y las *Deep Boltzmann Machines (DBM*).

Las *Boltzmann Machines* (BM), de forma muy general, son una especie de redes neuronales con conexiones bidireccionales y que dividen sus unidades en dos grupos: visibles o invisibles, conectadas todas con todas. Durante el entrenamiento, modifican sus pesos hasta encontrar el estado de más baja energía posible. Esta energía es similar a una función de coste. No tienen unidades de salida, sino que modifican la propia entrada en la búsqueda de ese estado de baja energía (Ilustración 3).

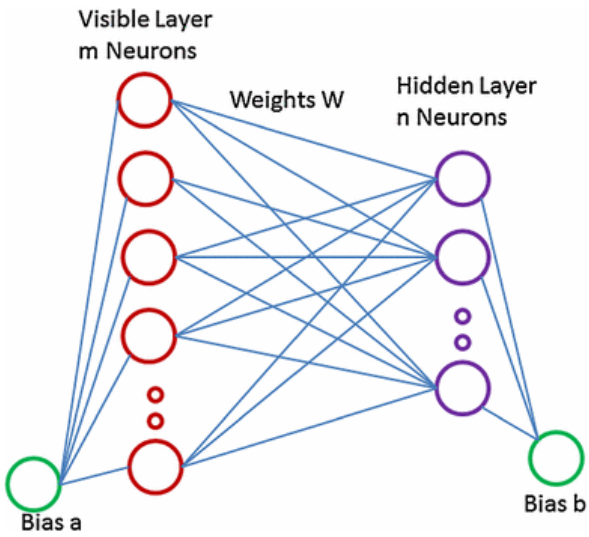

Ilustración 3. Restricted Boltzmann Machine (Chopra y Yadav, 2017)

Las RBMs son un caso especial de *Boltzmann Machines* que separan las unidades visibles y las unidades invisibles en dos capas, de tal manera que las unidades de una misma capa no se conectan entre sí. Minimizando la función de energía, estos sistemas reconstruyen el propio *input* x estimando la distribución de probabilidad de este (Salakhutdinov, Mnih, y Hinton, 2007).

Se basan en la idea de que la probabilidad de que un sistema esté en un determinado estado es inversamente proporcional a la energía de ese estado. Es decir, que el sistema tiende a buscar un estado de baja energía, razón por la cual se minimiza la función anterior.

La capa oculta h contendrá una representación de características que es capaz de reconstruir un *input* de forma aproximada. Por tanto, puede hacerse uso de la información de esta capa de forma similar a la capa oculta de un autoencoder. Li et al. (2015) crean una red neuronal MLP (*Multi-layer Perceptron)* en la que cada capa se inicializa como una RBM, para clasificar pacientes con Alzheimer o con deterioro cognitivo leve.

En la Ilustración 4 se muestra la estructura de una DBM:

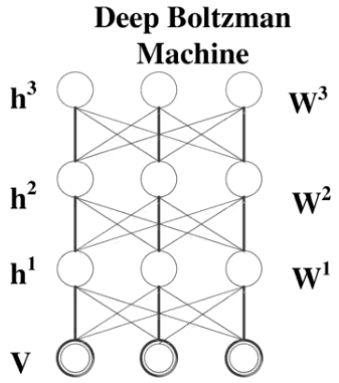

Ilustración 4. Deep Boltzmann Machine (Salakhutdinov y Hinton, 2009)

Las RBMs se pueden anidar para formar las DBMs (*Deep Boltzmann Machines*) (Salakhutdinov y Hinton, 2009), de forma similar a como los *stacked autoencoders* anidaban autoencoders básicos. De esta forma, cada capa es una RBM y se utiliza entrenamiento voraz por capas (Bengio et al., 2007) para ir obteniendo representaciones de mayor nivel conforme aumenta la profundidad.

H.-I. Suk et al. (2014) hacen uso de una DBM para obtener características representativas de los datos de entrada. Posteriormente, crean un sistema para fusionar la información extraída de diferentes modalidades (IRM y TEP), y con ello llevar a cabo una clasificación de pacientes con Alzheimer mediante una SVM.

CAPAS CONVOLUCIONALES

Los modelos básicos anteriores de aprendizaje no supervisado (autoencoders y RBM) pueden combinarse para inicializar capas sucesivas de una red neuronal. Esto es muy útil cuando se dispone de un conjunto de datos de proporciones limitadas, pues se puede llevar a cabo un pre-entrenamiento no supervisado que permita una convergencia más rápida durante el posterior entrenamiento supervisado.

En general, para construir capas convolucionales, los AE y SAE suelen ser bastante más indicados, pues permiten obtener fácilmente filtros convolucionales. El problema es que, aunque estos modelos pueden adaptarse para trabajar con dos o más dimensiones, en realidad funcionan de forma unidimensional (Masci, Meier, Cire\csan, y Schmidhuber, 2011).

Por suerte, existe una variante de estos métodos que trabaja nativamente en dos o más dimensiones, y que se han utilizado en algunos de los mejores modelos de diagnóstico del Alzheimer: los autoencoders convolucionales o CAE (Masci et al., 2011). Se trata de autoencoders que hacen uso de capas convolucionales para codificar el *input*, y luego llevan a cabo la operación inversa para decodificar el mapa de características resultante y reconstruir de nuevo el *input*.

Son modelos bastante más complejos que los anteriores, pues pueden incluso hacer uso de capas *fully connected* (todas las unidades de una capa conectadas con todas las de la siguiente), o capas de *pooling*. En la Ilustración 5 se muestra un ejemplo.

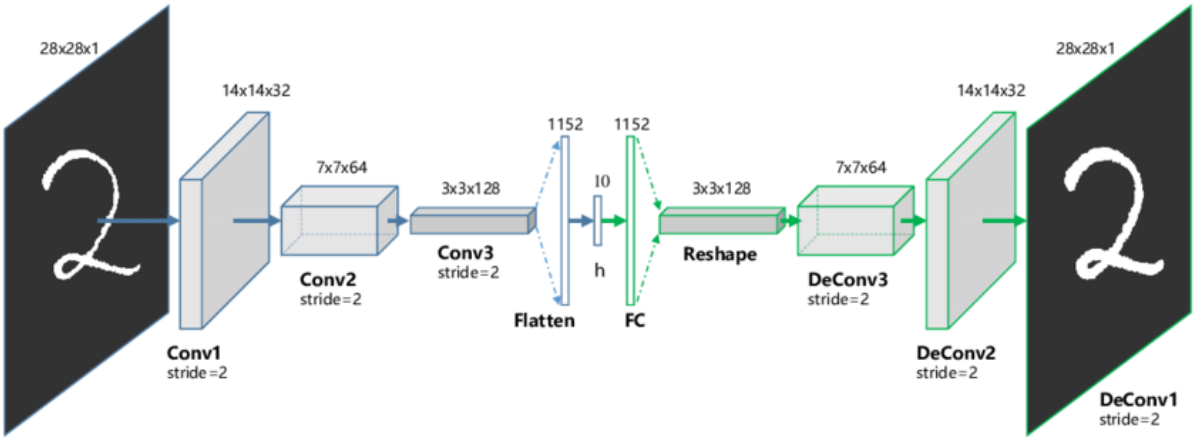

Ilustración 5. Autoencoder convolucional (Guo, Liu, Zhu, y Yin, 2017)

Estos modelos pueden emplearse para realizar una extracción de características profunda de imágenes médicas (M. Chen, Shi, Zhang, Wu, y Guizani, 2017). En el caso concreto del diagnóstico del Alzheimer, se han utilizado de forma anidada para construir algunos de los mejores modelos del estado del arte actual (Hosseini-Asl, Gimel'farb, et al., 2016; Hosseini-Asl, Keynton, et al., 2016), inicializando CNN 3D profundas.

Sin embargo, estas ideas de pre-entrenamiento por medio de extracción de características con autoencoders están, poco a poco, cayendo en desuso. Cada vez es más complicado apreciar mejoras de rendimiento con esta técnica, probablemente debido a técnicas como *Batch Normalization* o al uso predominante de unidades ReLU, que permiten entrenar modelos supervisados con mucho éxito (Hellström, 2017). La única razón de peso para hacer uso de esta técnica sería si se dispusiera de un gran conjunto de datos no etiquetado

y pocos datos etiquetados. De hecho, después del trabajo de Hosseini-Asl, Gimel'farb, et al. (2016), posteriores investigaciones han ignorado esta técnica.

2.4.2 *Convolutional Neural Networks*

Los modelos explicados hasta el momento han sido muy importantes durante años en el análisis de imágenes médicas en general. Sin embargo, en los últimos años, se aprecia un claro auge en el uso de las redes neuronales convolucionales (LeCun et al., 1998). En el caso del diagnóstico del Alzheimer, han explotado de forma más directa en los últimos tres años, con diferentes publicaciones que han logrado resultados excelentes, y siendo además modelos mucho más sencillos que los de años anteriores (Ding et al., 2018; Korolev, Safiullin, Belyaev, y Dodonova, 2017; Sarraf et al., 2016).

Esto no quiere decir que no se utilizasen anteriormente. De hecho, los primeros acercamientos comenzaron a apreciarse poco después de la publicación de Krizhevsky et al. (2012), que marcó un punto de inflexión en el uso de estos modelos para visión por computador. Sin embargo, estos usos de las CNN venían acompañados por otros modelos, principalmente los vistos en el apartado 2.4.1 para extracción de características. Lo que se está logrando en los últimos años es hacer uso directa y únicamente de las CNN para resolver este tipo de problemas.

Las redes neuronales convolucionales son similares a las redes neuronales tradicionales, pero añaden una serie de elementos pensados para el tratamiento de imagen:

- Las capas convolucionales consisten en utilizar un filtro convolucional que recorra la imagen de entrada, produciendo un valor por cada píxel sobre el que se aplica el filtro. En estas capas también se aprenden pesos y *biases* como en las capas corrientes.

- Las capas *pooling* permiten reducir las dimensiones tratadas y proporcionan invariancia a la rotación. Consisten en utilizar una ventana deslizable, como el filtro convolucional, pero escogiendo siempre el valor máximo por ventana (*max pooling*). De esta manera, se reduce la cantidad de información. En ciertas redes, se calcula la media de los valores comprendidos por la ventana (*mean pooling*), pero esto cada vez se utiliza menos porque da peores resultados.

En la última década, se han dado una serie de avances importantes en redes convolucionales que han permitido que, poco a poco, se hayan acercado al análisis de

imágenes médicas. A continuación, se resumen los principales hitos que deberían tenerse en cuenta:

1. AlexNet (Krizhevsky et al., 2012). Esta red marcó un punto de inflexión en la historia de la visión por computador. Hizo uso por primera vez de diversas técnicas modernas, como las unidades ReLU, lo cual daba mucho mejores resultados que la tradicional sigmoide, o el entrenamiento con GPUs.

2. VGGNet (Simonyan y Zisserman, 2014). Fue una red que aumentó considerablemente la profundidad, llegando a tener hasta 19 capas. Utilizaba filtros convolucionales más pequeños, de 3x3 píxeles. Al contrario que con AlexNet, los autores publicaron sus dos mejores modelos, permitiendo a la comunidad acceder a los mismos. Es decir, que hicieron posible realizar *fine-tuning* de una red puntera, pre-entrenada en *ImageNet*.

3. Inception (Szegedy et al., 2015). La red *GoogLeNet* hacía uso de la arquitectura Inception, que se caracterizaba por mantener la complejidad con respecto a su competencia, pero aumentando su profundidad y anchura, llegando a las 22 capas. Los autores buscaron mejorar la eficiencia del entrenamiento, con la intención de hacer factible la aplicación de estos modelos en entornos reales, donde los *smartphones* y los sistemas empotrados tenían una gran presencia.

4. *DeepImage* (Wu, Yan, Shan, Dang, y Sun, 2015). En este caso, los autores se centraron en desarrollar el hardware más adecuado posible, y hacer uso de un software muy bien optimizado para ese hardware en concreto. Demostraron cómo la aumentación de datos realizada de forma agresiva puede ayudar a reducir el sobreajuste.

5. ResNet (He, Zhang, Ren, y Sun, 2015). A estas alturas, ya se había comprobado que, teniendo suficientes datos, aumentar el tamaño de la red siempre otorgaba mejoras. Así, en Microsoft presentaron un nuevo concepto, los bloques residuales (Ilustración 6), que simplificaban el entrenamiento de redes muy profundas. Crearon una red residual de 152 capas, 8 veces más profunda que VGGNet, y pese a ello menos costosa de entrenar.

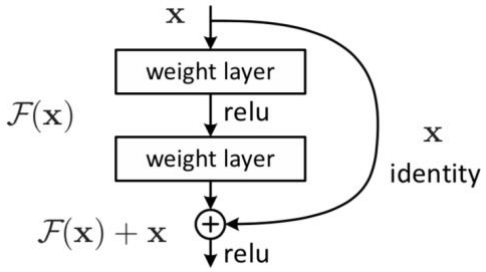

Ilustración 6. Bloque residual de una ResNet (He, Zhang, Ren, y Sun, 2016)

6. InceptionV3 (Szegedy, Vanhoucke, Ioffe, Shlens, y Wojna, 2016). Se trataba de una evolución de la arquitectura Inception original, que seguía la misma filosofía: darle mucha importancia a la eficiencia para poder desplegar el modelo en un mayor número de entornos.

Todas estas redes eran entrenadas con Imagenet y cada una, en su momento, fue la ganadora de la competición ILSVRC («Large Scale Visual Recognition Challenge», 2015). Al final, las arquitecturas ResNet e InceptionV3 son las más utilizadas. De hecho, en la mayoría de las librerías de programación de redes neuronales, como Keras (Chollet y otros, 2015), están publicados ambos modelos, ya entrenados con Imagenet. Esto permite realizar reajuste o *fine-tuning* en el área de aplicación que se desee, y el análisis de imágenes médicas no es excepción.

Esta última técnica, el *fine-tuning*, consiste en aprovechar los pesos de una red neuronal ya entrenada y, utilizando aprendizaje supervisado, terminar de ajustarlos a un área de aplicación en concreto. De esta manera, se aprovecha una red puntera y se adapta a la tarea deseada. Es una técnica muy utilizada hoy en día.

Idealmente, la red debería estar entrenada con imágenes similares a las del área de aplicación. Por ejemplo, para el diagnóstico del Alzheimer, se podría utilizar una red ya entrenada con un conjunto de datos IRM empleado para estimar la cantidad de materia gris. Sin embargo, esto muchas veces no es posible, pues la cantidad de datos disponible en análisis de imagen médica suele estar limitado (apartado 2.6.1).

Por ello, aunque las imágenes son muy diferentes, se tiene que hacer uso de redes pre-entrenadas con Imagenet, pues es un conjunto de imágenes muy grande y muy variado, con 1000 clases diferentes. Por un lado, esto es bueno porque las imágenes de esta base de datos varían mucho en comparación con las MRI. Por otro lado, los detalles son más evidentes en una imagen en la que aparece un perro que en una imagen donde se supone que hay indicadores de Alzheimer. No obstante, esto último es un inconveniente mucho

mayor a ojos de un humano que para una CNN, que es capaz de extraer patrones mucho más sutiles.

Así pues, las redes InceptionV3 y ResNet son comúnmente reajustadas para aplicarlas en análisis de imágenes médicas. La Tabla 3 incluye alguna de las publicaciones más relevantes al respecto. Más allá del diagnóstico del Alzheimer, han servido para clasificar exitosamente más de 2000 variantes de cáncer de piel (Esteva et al., 2017), han logrado detectar retinopatía diabética (Gulshan et al., 2016), y han permitido realizar el diagnóstico de cáncer de mama con muy buenos resultados (Vesal et al., 2017). Las dos primeras publicaciones hacían uso exclusivamente de InceptionV3, mientras que la última también la compara con una ResNet50 (red residual con 50 capas).

Tabla 3.
Publicaciones que hacen uso de redes neuronales convolucionales.

PUBLICACIÓN	ENFERMEDAD OBJETO DE DIAGNÓSTICO
(Sarraf et al., 2016)	Alzheimer
(Gulshan et al., 2016)	Retinopatía diabética
(Korolev et al., 2017)	Alzheimer
(Esteva et al., 2017)	Cáncer de piel (más de 2000 variantes)
(Vesal et al., 2017)	Cáncer de mama
(Ding et al., 2018)	Alzheimer

Con respecto al diagnóstico del Alzheimer, el primer trabajo relevante totalmente enfocado en el uso de CNN sería probablemente el de Sarraf et al. (2016). Transformando imágenes MRI y fMRI a formato PNG, entrenaron dos redes neuronales: una basada en la arquitectura LeNet-5 (LeCun et al., 1998) y otra basada en GoogLeNet Inception V1 (Szegedy et al., 2015). Se encontraron con que la primera convergía más rápidamente, aunque la segunda lograba resultados ligeramente mejores en la mayoría de los casos. También construyeron un algoritmo de toma de decisiones, basado en la combinación de ambos modelos y otros criterios, con el cual rozaron el 100% de exactitud.

Posteriormente, Korolev et al. (2017) buscaron demostrar que podía hacerse uso de CNN de forma sencilla para generar características de forma totalmente automática y clasificar. Para ello, simplifican al máximo la creación de modelos, utilizando únicamente 231 imágenes. Construyen una CNN similar a VGGNet (Simonyan y Zisserman, 2014), que llamaron VoxCNN; y lo comparan con un modelo VoxResNet (H. Chen, Dou, Yu, y Heng, 2016), el

cual es básicamente un modelo ResNet (He et al., 2015) adaptado para imágenes 3D. Comparándose con Hosseini-Asl, Gimel'farb, et al. (2016), obtuvieron resultados notablemente peores, pero destacaban la sencillez a la hora de implementar su modelo.

Por último, Ding et al. (2018) hacen *fine-tuning* de una red InceptionV3 utilizando imágenes 18F-FDG PET, utilizando la librería Keras (Chollet y otros, 2015) para implementar los modelos y SciPy (Jones, Oliphant, Peterson, y otros, 2001) para preprocesar las imágenes. Realizando una comparación directa con radiólogos, demostraron que su modelo los superaba de forma estadísticamente significativa, especialmente a la hora de anticipar la aparición de Alzheimer con más de seis años de antelación.

2.5 Técnicas de preprocesamiento

Una fase fundamental en la construcción de un modelo de aprendizaje automático es el preprocesamiento de los datos, en el que se busca prepararlos para que puedan ser aprovechados de la mejor forma posible por los modelos. Por ejemplo, en la construcción de la red AlexNet, se ajustan todas las imágenes a una dimensiones de 256x256 (Krizhevsky et al., 2012). En VGGNet, se resta a cada valor de píxel la media del total (Simonyan & Zisserman, 2014).

De la misma forma, existen una serie de técnicas de preprocesamiento de imágenes médicas que son comunes a una amplia variedad de investigaciones sobre el diagnóstico del Alzheimer. En apartados anteriores, se explicó cómo las publicaciones de la pasada década hacían uso de características definidas de forma manual, lo cual requería de técnicas de preprocesamiento muy complejas. Sin embargo, el uso de redes convolucionales y autoencoders para la extracción automática de características simplifica ligeramente este proceso. Al final, se deben destacar principalmente dos procedimientos clave: el registro de imágenes médicas y la extracción del cráneo.

El registro consiste en adaptar una imagen determinada a otra de referencia, que llamamos atlas, buscando que los píxeles de ambas representen las mismas estructuras anatómicas (Klein et al., 2009; Woods, Mazziotta, Cherry, y otros, 1993). De esta forma, es más sencillo que una CNN identifique como relevante una determinada región de las imágenes, pues en todas ellas se estaría representando la misma información. Existe una gran variedad de algoritmos de registro diferentes (Klein et al., 2009), aplicables no sólo a neuroimagen (Hosseini-Asl, Keynton, et al., 2016; Sarraf et al., 2016; H.-I. Suk et al., 2014, 2015) sino también a otros campos de la medicina, como el cáncer de mama (Rueckert et al., 1999).

La extracción del cráneo (en inglés, *skull stripping*) consiste, como su propio nombre indica, en eliminar la información del cráneo que aparece en las imágenes de resonancia magnética. El objetivo es obtener una imagen final lo más limpia posible, y que contenga únicamente la información relevante para la tarea en curso. En el caso del Alzheimer, ya se expusieron en la sección 2.3 los biomarcadores más relevantes, y evidentemente ninguno de estos se encuentra en el cráneo. Por ello, ciertas publicaciones utilizan algoritmos de extracción del cráneo y otras regiones no cerebrales (Sarraf et al., 2016; H.-I. Suk et al., 2014, 2015), o directamente hacen uso de un conjunto de datos de imágenes con el cráneo ya extraído (Korolev et al., 2017).

En la Ilustración 7 se puede observar cómo el cráneo forma parte importante de la información representada en una imagen de resonancia magnética:

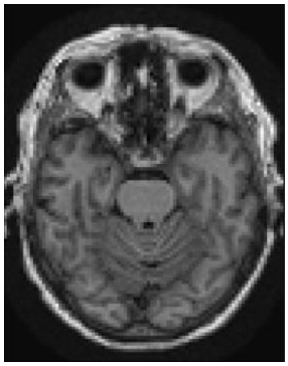

Ilustración 7. Plano axial bajo de una IRM sin el cráneo extraído

Más allá de estos dos procedimientos, se encuentran otras técnicas más genéricas, como la normalización de las imágenes. Esta puede descomponerse en dos variantes: la normalización de la intensidad y la normalización espacial. La primera de estas se basa en adaptar el rango de valores de los píxeles según un determinado criterio, como puede ser reducirlos al rango $[0,1]$ o $[-1,1]$ o restando la media y dividiendo por la desviación típica, lo que se conoce como *whitening* (Rajchl, Ktena, y Pawlowski, 2018). La segunda, por su parte, consiste en adaptar los píxeles (o vóxeles en 3D) para que representen un determinado espacio (o volumen) (Rajchl et al., 2018). Por ejemplo, Ding et al. (2018) adaptan las imágenes 3D para que cada vóxel represente $2mm^3$ de espacio.

En este sentido, el registro de imágenes puede considerarse una forma de normalización espacial (Rajchl et al., 2018).

2.6 Principales retos

Los distintos usos de la inteligencia artificial en aplicaciones clínicas se enfrentan a una serie de retos, algunos de ellos similares a los que se encuentran estas mismas técnicas en otros ámbitos. La gran mayoría son relativos a los datos, los cuales son la base fundamental para construir sistemas precisos y seguros. No obstante, también se presentan retos o dilemas éticos y filosóficos, así como otros problemas menores que pueden dificultar ciertas investigaciones.

2.6.1 Los datos

DISPONIBILIDAD

En el apartado 2.2.1, se explica cómo el aumento en el uso de los ECE, aproximándose al *Big Data*, ha sido un factor clave en el auge del aprendizaje automático en aplicaciones médicas. Si bien esto es completamente cierto, también hay que tener en cuenta que se está muy lejos de tener la información necesaria para lograr avances similares a los que se están obteniendo con imágenes naturales.

Póngase como ejemplo el trabajo de Krizhevsky et al. (2012), como podría ser cualquiera de los posteriores grandes avances en redes neuronales. Estas redes se suelen entrenar con las imágenes de la base de datos de Imagenet («Imagenet», 2016), que hoy dispone de más de 14 millones de imágenes. Si bien es cierto que estas se pueden clasificar en 1000 clases diferentes, la diferencia en la cantidad de datos anotados disponibles es muy grande si se compara con las bases de datos de información médica. Por ejemplo, en la base de datos OASIS-3 («OASIS Brains Datasets», s. f.), se dispone de imágenes MRI y PET de tan solo 1098 sujetos.

La cantidad de datos necesaria para acometer diferentes problemas de clasificación está aún pendiente de investigación (Greenspan et al., 2016). Si se presta atención a las últimas publicaciones exitosas, más allá del diagnóstico del Alzheimer, se encuentran modelos como los de Esteva et al. (2017) y Gulshan et al. (2016), que utilizan más de 128000 imágenes cada uno. Sin embargo, los resultados de Ding et al. (2018), que, como los anteriores, hace *fine-tuning* de una red InceptionV3 (Szegedy, Vanhoucke, et al., 2016) pre-entrenada en Imagenet para diagnosticar el Alzheimer, hace uso de poco más de 2000 estudios y obtiene peores resultados.

Muchas veces esto no es un problema de disponibilidad de datos en general, sino de su etiquetado. Sin etiquetas, no se puede llevar a cabo aprendizaje supervisado. En ocasiones,

se ha tratado incluso de externalizar el etiquetado de los datos (*crowdsourcing*) (Greenspan et al., 2016).

Por tanto, aunque los datos etiquetados disponibles han aumentado considerablemente en los últimos años, estos siguen siendo algo limitados en ciertos escenarios. En problemas complejos, esto puede llevar fácilmente a sobreajustar el modelo a los datos (*overfitting*), por lo que diversas soluciones se han propuesto. La más común es la extracción de múltiples parches (*patches*) aleatorios de las imágenes. Gupta, Ayhan, y Maida (2013), y posteriormente Payan y Montana (2015), hacen uso de esta técnica. Suk, Lee, Shen, Initiative, y otros (2014) llegan incluso a defender que hacer uso de parches se asemeja al procedimiento que siguen los radiólogos, examinando imágenes por regiones.

El problema es que no todos los parches extraídos de una imagen de la clase positiva tendrán características representativas de ésta. Como solución, en Tong et al. (2014) proponen todo un sistema de grafos para extracción de parches sin información redundante, y representando las relaciones entre los que se extraen de una misma imagen. De esta forma, los parches integran información contextual.

Otra solución, algo menos común que la anterior, es hacer uso de técnicas de aumento de datos o *data augmentation*. Su uso está algo limitado porque, al estar añadiendo información hasta cierto punto redundante, existe riesgo de aumentar el *overfitting*, en lugar de disminuirlo. Ding et al. (2018) realiza desplazamientos aleatorios en anchura y altura, así como *zoom* para aumentar la cantidad de datos de entrenamiento.

Sin embargo, la gran ventaja de su modelo probablemente sea que hace uso de *transfer learning*. Como ya se ha comentado, se trata de obtener una red neuronal previamente entrenada y hacer reajuste (*fine-tuning*) para adaptarla al problema en cuestión. Y es que el *transfer learning* es otra forma de hacer frente a las limitaciones en la cantidad de datos, pues la red ya viene entrenada y sólo es necesario reajustarla para adaptarla al área de aplicación, reduciendo la cantidad de datos necesarios para converger.

También se han dedicado investigaciones enteras a evitar el *overfitting* buscando directamente la manera de obtener más datos de entrenamiento, pero de forma auténtica, es decir, creando imágenes totalmente nuevas. El trabajo de Castro, Ulloa, Plis, Turner, y Calhoun (2015) logra resultados medianamente satisfactorios. Este interés por la generación de imágenes sintéticas se dispara tras la publicación de las redes generativas adversarias o GANs (Goodfellow et al., 2014). Están logrando resultados muy destacables en la generación de imágenes naturales sintéticas, lo cual las hace muy prometedoras en áreas

como la medicina. Otros modelos que podrían servir para este propósito serían los autoencoders variacionales (en inglés, *variational autoencoders*) o VAE (Ker et al., 2018).

Por último, existen otras técnicas para reducir el *overfitting*, más básicas, pero no por ello menos importantes. Hablamos de técnicas de regularización comunes, como *dropout*, hacer uso de *batch normalization*, utilizar unidades ReLU… entre otras.

DESBALANCEO DE CLASES

Otra dificultad a la que se enfrentan los investigadores es la baja representación de la clase positiva en los conjuntos de datos. Como es lógico, suele ser mucho más sencillo encontrar información de pacientes no enfermos que de pacientes enfermos, porque además estos tienen cierta tendencia a no compartir su información (apartado 2.6.2). Este problema se acentúa en el caso de las enfermedades raras. Para empeorar las cosas, la clase negativa suele estar fuertemente correlada, mientras que existe muchísima variación en la clase positiva (Greenspan et al., 2016).

Mazurowski et al. (2008) comprobaron en sus experimentos que el desbalanceo de clases afecta negativamente y de forma clara, incluso para pequeños desbalanceos. Llegaron a la conclusión de que eliminar observaciones de la clase sobre-representada (*undersampling*) casi nunca es bueno, mientras que, en ocasiones, duplicar observaciones de la clase infrarrepresentada (*oversampling*) puede mejorar los resultados muy ligeramente. Sin embargo, al ser tan pequeña la mejora, el *oversampling* no ha llegado a cobrar demasiada importancia en la bibliografía.

Para atacar este problema, otros investigadores fuerzan el balanceo de los datos, es decir, seleccionan el mismo número de instancias de cada una de las clases (Gupta et al., 2013; Klöppel et al., 2008; Payan y Montana, 2015). Pero, en general, la mayoría de las publicaciones no le prestan mucha atención a esto, aunque también es cierto que en ninguna se da un desbalanceo demasiado notable.

VARIEDAD ESTRUCTURAL

La variedad estructural presente en las imágenes médicas, así como las diferentes formas de tratar esa estructura, es un reto al que la mayoría de las publicaciones dedican mucha importancia. Cuando se habla de variedad estructural, se refiere a la variedad de tipos de imágenes médicas que pueden obtenerse. Ya se mencionaron al comienzo de esta sección la gran variedad de escaneos que pueden llevarse a cabo del cerebro, y cada uno de ellos tiene su propia estructura.

En general, las imágenes MRI estructurales son las más utilizadas para el diagnóstico CADx del Alzheimer. Están presentes en la totalidad de la literatura, excepto en el trabajo de Ding et al. (2018), que hace uso exclusivamente de imágenes 18F-FDG PET. También es bastante común combinar imágenes MRI estructurales con imágenes PET, e incluso hay ciertos investigadores que las han combinado con medidas de líquido cefalorraquídeo (CSF – *Cerebrospinal fluid*) (H.-I. Suk et al., 2015; H. Il Suk y Shen, 2013). El tipo de escaneo menos utilizado sería el MRI funcional (fMRI), que tan sólo aparece combinado con MRI estructural en una ocasión (Sarraf et al., 2016).

También hay diferentes formas de tratar estas estructuras. Se conoce que muchas imágenes médicas, como las del cerebro, están en tres dimensiones. Así pues, es de suponer que el tratamiento 3D es el más indicado para no perder información. Sin embargo, aunque las redes 3D se utilizan bastante, muchos investigadores preprocesan las imágenes para llevarlas a 2D. De hecho, esto ocurre en la mayoría de los casos.

Payan y Montana (2015) presentaron el uso de filtros convolucionales en 3D como una novedad, aunque estos eran aprendidos a partir de un autoencoder disperso (*sparse*). Estos consideran el valor de cada vóxel por igual, como parte de la misma dimensión. Hosseini-Asl, Gimel'farb, y El-Baz (2016) y Hosseini-Asl, Keynton, y El-Baz (2016) van un paso más allá haciendo uso de autoencoders convolucionales. Más recientemente, Korolev, Safiullin, Belyaev, y Dodonova (2017) utilizan las imágenes 3D directamente sobre una red neuronal convolucional 3D, ofreciendo un enfoque aún más directo, y a la vez más sencillo.

Más allá de estos trabajos, la tendencia general ha sido hacer uso de 2D, utilizando alguna técnica de preprocesamiento o escogiendo múltiples planos de cada imagen 3D. Ciertos trabajos llegan incluso a transformar las imágenes a formato PNG (Sarraf et al., 2016). También es muy destacable el procedimiento seguido por Ding et al. (2018), en el que escogen 16 planos de las imágenes 3D y los colocan en una imagen 2D formando una matriz 4x4, con fondo negro.

Esta tendencia se debe principalmente a que es computacionalmente mucho más eficiente tratar imágenes 2D. Para visualizar hasta qué punto, se puede pensar que, por cada valor de profundidad de la tercera dimensión, se está procesando una imagen extra. Esto en imágenes RGB significaría procesar el triple, pero con dimensiones 68×95×79, el incremento en la complejidad es mucho mayor.

Las dimensiones mencionadas como ejemplo son las que se utilizan en la publicación de Payan y Montana (2015). Estos también llevaron a cabo un experimento para comprobar si su red 3D tenía una mejora de rendimiento significativa con respecto a una red 2D similar a

la de Gupta et al. (2013). La mejora era de en torno al 4% de exactitud (*accuracy*) en el mejor caso, pero en la clasificación binaria AD vs HC (enfermo de Alzheimer vs paciente sano), la exactitud era exactamente la misma.

COMBINACIÓN DE TIPOS

Una idea que tiene cierta presencia en la bibliografía es acompañar la información contenida en las imágenes médicas con otros campos categóricos o numéricos. Por ejemplo, en un sistema de diagnóstico del Alzheimer, se podría combinar una imagen MRI con la edad de la persona, pues una persona mayor tiene más probabilidades de sufrir la enfermedad.

Esto, en la práctica, no termina de funcionar del todo bien. Aunque en algunos casos puede haber una mejora, la cantidad de información presente en las imágenes es tan grande en comparación con los campos simples, que estos últimos se vuelven irrelevantes (Litjens et al., 2017). Las redes profundas tienden a centrarse en las características extraídas de las imágenes, otorgando pesos mucho menores a la información adicional.

Otra forma de combinar información con las imágenes es por medio de informes completos, no sólo campos individuales. Aquí el problema anterior no se da, pues un informe puede incluir mucha información, que al estar etiquetada puede mejorar los resultados (Litjens et al., 2017). Ahora bien, hacer esto implicaría redactar el informe de antemano, lo cual exige mucho trabajo y muchas pruebas, deshaciendo un poco la ventaja que se buscaba en un inicio: simplificar el trabajo al radiólogo lo máximo posible.

2.6.2 Cuestiones éticas y filosóficas

Los ECE representan información muy sensible sobre los pacientes. Esta es una realidad con la que los médicos tratan diariamente, siendo fundamental el secreto profesional doctor-paciente. Además, con el creciente debate en los medios acerca de la importancia de la privacidad de los datos de las personas, han surgido leyes como la GDPR (*General Data Protection Regulation*) en la Unión Europea, que acarrea severas multas si no se manejan adecuadamente los datos privados, otorgando un mayor nivel de control a las personas sobre cómo su información es tratada por empresas e instituciones («2018 reform of EU data protection rules», 2018).

Todo esto, como es evidente, limita los conjuntos de datos que pueden reunirse. Para construir bases de datos de imágenes médicas, se debe obtener el consentimiento de las personas cuya información se está compartiendo. Estas condiciones acentúan el problema del desbalanceo de clases mencionado en el apartado anterior, pues una persona enferma estará menos predispuesta a compartir su historial clínico.

Otra cuestión bastante importante es la confianza. En concreto, la confianza hacia los sistemas de inteligencia artificial. Hoy en día, todavía existe un gran desconocimiento en la población acerca de lo que es o cómo funciona realmente la inteligencia artificial. Los escenarios apocalípticos presentados en el cine y otros medios, así como las declaraciones de ciertos personajes públicos, acentúan esta desconfianza. Como resultado, buena parte de la población es reacia a la incursión de estos sistemas en aspectos relevantes de la vida, como lo es la medicina.

Esta desconfianza no procede únicamente de los pacientes, sino también de los propios médicos, quienes también pueden estar afectados por ese desconocimiento. Sin embargo, incluso para aquellos que logran entender las bases reales de la Inteligencia Artificial o que, sencillamente, no desconfían, es complicado incluir estos sistemas en su trabajo diario. Esto se debe a que, hoy en día, no se dispone de los medios para explicar las decisiones que toman los modelos.

El diagnóstico de una persona es un tema especialmente sensible, que puede acarrear consecuencias muy graves si se cometen errores. Por ello, un médico no puede tomar una decisión al respecto a la ligera, simplemente confiando en los resultados obtenidos por un modelo. Necesita una explicación clara de por qué se toma la decisión que se toma, especialmente cuando el diagnóstico realizado por el médico entra en conflicto con el resultado determinado por el modelo.

Por todo esto, lo más sensato probablemente sería continuar delegando la decisión final en el propio médico y no en el modelo, aunque esto retrase la adopción de la Inteligencia Artificial en el diagnóstico. Utilizar un *ensemble* de diferentes modelos también podría ser una buena idea.

Este es un problema recurrente en el despliegue de modelos de Inteligencia Artificial en entornos reales. Surge como resultado de que los modelos que mejor funcionan hoy en día son modelos de caja negra, es decir, modelos cuyos resultados o procedimientos internos son extremadamente complicados o directamente imposibles de interpretar.

El primero de estos problemas podría solucionarse con el tiempo, según la Inteligencia Artificial continúa penetrando en aspectos más cotidianos de la vida humana. Por otro lado, el problema de la caja negra es un campo de investigación abierto y bastante activo (Montavon, Lapuschkin, Binder, Samek, y Müller, 2017; Zeiler y Fergus, 2014), aunque apenas tratado en el estado del arte del diagnóstico del Alzheimer. Tan sólo Ding et al. (2018) crea un sistema que computa los gradientes para la clase AD (enfermo de

Alzheimer), y crea un mapa de saliencia o *saliency map* que permite visualizar las áreas de las imágenes que han sido consideradas de importancia por la red neuronal.

2.6.3 Clasificación multiclase y otras limitaciones

En este apartado, se mencionan otros retos menores a los que se enfrentan los investigadores, comenzando por la clasificación multiclase. Los modelos que presentan los investigadores suelen dedicarse a una enfermedad en concreto. Por tanto, hacen uso de un conjunto de datos en el que la enfermedad que puede aparecer es únicamente la que están tratando. Como mucho, tienen que manejar las diferentes fases de una misma enfermedad. Por ejemplo, en el caso del Alzheimer, se suelen manejar tres clases: AD, MCI y HC.

Sin embargo, estos enfoques simplifican excesivamente el problema real, que es el diagnóstico de un paciente. Cuando se percibe cierto deterioro cognitivo, múltiples tipos de demencia pueden ser enfermedades candidatas, no sólo el Alzheimer. Por ello, pensando en la puesta en producción de uno de estos sistemas, sería necesario entrenar un modelo para que detecte un mayor abanico de enfermedades. En este contexto, nos encontraríamos de nuevo con el problema de la cantidad limitada de datos etiquetados, pues sería necesario un incremento significativo de éstos por cada nueva clase o enfermedad.

Un buen ejemplo de implementación de un modelo completo sería la publicación de Esteva et al. (2017). Tratan más de 2000 variantes diferentes de cáncer de piel, logrando resultados comparables a los obtenidos por dermatólogos expertos. Eso sí, disponen de un conjunto de datos de más de 129000 imágenes.

Otras dificultades importantes serían las limitaciones de hardware y la optimización de hiperparámetros. La primera es especialmente importante cuando se tratan imágenes 3D. La segunda no suele mencionarse en la bibliografía, aunque lo más inteligente sería hacer uso de la búsqueda aleatoria de hiperparámetros (Bergstra y Bengio, 2012), la cual es la que mejores resultados suele otorgar en casi cualquier ámbito. Más recientemente, investigadores de Google propusieron un método basado en algoritmos genéticos para optimización de hiperparámetros (Real, Aggarwal, Huang, y Le, 2018), pero hasta el momento no se ha utilizado en aplicaciones médicas y su coste computacional es demasiado alto.

2.7 Conclusiones de la revisión

Con toda la información expuesta hasta el momento, la conclusión principal que se puede extraer es que no existe debate acerca de cuál es el tipo de modelo de aprendizaje automático que debería utilizarse. En su lugar, la discusión se centra en cómo procesar los datos y cómo llevar a cabo el entrenamiento del modelo, el cual es una red neuronal convolucional. Las máquinas de vector de soporte han perdido completamente el protagonismo en los últimos cinco años.

Con respecto al preprocesamiento, lo ideal es registrar las imágenes disponibles a unas dimensiones comunes (lo que se conoce como un atlas) para que los modelos puedan centrarse en determinadas regiones de las imágenes, teniendo la seguridad de que en todas ellas se representa un mismo elemento anatómico. Además de esto, la información de interés a la hora de diagnosticar el Alzheimer se encuentra en el tejido cerebral, haciendo que las regiones correspondientes al cráneo sean irrelevantes. Estas regiones son capturadas en las imágenes IRM. Por esta razón, la extracción del cráneo se vuelve una práctica muy común, con el objetivo de entrenar el modelo únicamente con la información relevante. En modelos 2D, también es muy importante el procedimiento de conversión desde las imágenes 3D originales.

Con respecto a los modelos, existen variaciones en la forma de entrenar las redes neuronales convolucionales:

- Realizar reajuste de redes pre-entrenadas en Imagenet. En este caso, las imágenes tienen que convertirse a dos dimensiones, pues estas redes están preparadas para trabajar con imágenes RGB (bidimensionales con un tercer canal para el color). En este punto, también se puede hacer uso de diferentes arquitecturas, como Inception o ResNet.

- Utilizar redes de diseño propio o con pesos inicializados con técnicas tradicionales. Las redes pueden trabajar con información en 2D o en 3D, dependiendo de si se prefiere disminuir el coste del entrenamiento o mantener el detalle, respectivamente.

- Utilizar redes de diseño propio, pero llevando a cabo un pre-entrenamiento no supervisado. Se trata de construir una red convolucional a partir de modelos no supervisados, como los autoencoders. Esta opción está cayendo en desuso.

Estas tres formas de diseñar y entrenar los modelos son las que han protagonizado la investigación en este campo en los últimos cinco años. Por ello, la fase descriptiva de este trabajo se va a centrar en estos métodos.

3 Comparativa experimental

En esta sección comienza la fase descriptiva de esta tesis. Se describe el procedimiento seguido para implementar los diferentes modelos, así como el preprocesamiento llevado a cabo sobre los datos. También se analizan los resultados y se comparan entre sí, teniendo siempre en cuenta otras implementaciones llevadas a cabo en trabajos similares.

Para el desarrollo del código, se ha utilizado el lenguaje de programación Python y los cuadernos de Google Colaboratory («Colaboratory: Frequently asked questions», 2018). En la implementación, se han utilizado múltiples librerías diferentes, especialmente durante la fase de preprocesamiento de los datos. Con respecto a los modelos, se ha hecho uso de Keras (Chollet y otros, 2015) funcionando sobre Tensorflow (Abadi et al., 2015).

Otra opción era hacer uso de una librería dedicada al prototipado de sistemas de *Deep Learning* para el campo de la medicina: el *Deep Learning Tool Kit* (DLTK) (Pawlowski et al., 2017). Sin embargo, este funciona directamente sobre Tensorflow y requiere de la misma interfaz, mientras que Keras es mucho más sencillo de utilizar. Además, Tensorflow se encuentra en pleno proceso de migración hacia una nueva versión, la 2.0, donde adopta muchas de las ideas de Keras y modifica notablemente su comportamiento por defecto. Por estas razones, se tomó la decisión de no hacer uso directo de DLTK, aunque sí se toma inspiración de esta librería para la implementación de los modelos. Los detalles al respecto se especifican en los apartados dedicados a los propios modelos (apartados 3.2.1 y 3.2.2).

3.1 *Dataset*

Para el *Alzheimer's Disease Neuroimaging Initiative*. Los datos utilizados en la preparación de este artículo se obtuvieron de la base de datos del *Alzheimer's Disease Neuroimaging Initiative* (ADNI) (adni.loni.usc.edu). Como tal, los investigadores dentro del ADNI contribuyeron al diseño y la implementación del ADNI y/o proporcionaron los datos, pero no participaron en el análisis ni en la redacción de este informe. Una lista completa de los investigadores del ADNI se puede encontrar en:
http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

El ADNI surgió en 2003 como una asociación público-privada, liderada por el investigador principal Michael W. Weiner, MD. El objetivo principal del ADNI ha sido probar si la resonancia magnética (RM), la tomografía por emisión de positrones (TEP), otros marcadores biológicos y la evaluación clínica y neuropsicológica pueden combinarse para

medir la progresión del deterioro cognitivo leve (DCL) y la enfermedad de Alzheimer temprana. Para obtener información actualizada, consulte www.adni-info.org.

De los datos cedidos por el ADNI, se obtuvieron más de 3000 imágenes de resonancia magnética (IRM) en la modalidad T1 ponderada (T1-*weighted*). Los pacientes tenían una edad en un rango de entre 55 y 92 años, y se les asignaba una etiqueta de las tres posibles mencionadas en apartados anteriores: AD (enfermo de Alzheimer), MCI (deterioro cognitivo leve) y HC (paciente sano). Cabe mencionar que, en ADNI, la clase de paciente sano se denota por NC, que significa *normal cohort* o grupo normal (Korolev et al., 2017). En la *Tabla 4* se puede observar la distribución completa.

Tabla 4
Distribución de las clases para las imágenes RM

	Nº IRM
AD (Enfermo de Alzheimer)	636
MCI (Deterioro cognitivo leve)	1636
CN (Paciente sano)	903
TOTAL	3175

Los pacientes con imágenes RM tienen entre 55 y 92 años. El tamaño del conjunto de datos se encuentra en la línea de lo que se ha utilizado para las principales investigaciones referenciadas en la sección 2. También es importante destacar que el conjunto de datos está considerablemente desbalanceado.

Los datos utilizados se han limitado a imágenes, sin estar complementadas con ninguna otra información. En el apartado 1.4 se adelantó que la única información extra disponible era la edad, la cual es insignificante en comparación con la información que se puede extraer de las imágenes. En el apartado 2.6.1, se explicó cómo esto provoca que los modelos ignoren por completo la información relativa a la edad, razón por la cual se ha tomado la decisión de utilizar únicamente las imágenes para la experimentación.

De todos estos datos, se mantiene siempre un 15% al margen del entrenamiento, para usarlos como conjunto de prueba. Al mismo tiempo, del conjunto de entrenamiento restante, se separa otro 15% para utilizarlo como conjunto de validación e ir evaluando la calidad del modelo según avanza el entrenamiento (con cada *epoch*).

3.1.1 Preprocesamiento

Las imágenes de resonancia magnética obtenidas de la base de datos del ADNI tenían una enorme variedad de formas, y diferentes conjuntos habían sido pre-procesados de antemano con múltiples técnicas. Por ello, era necesario registrar las imágenes a un atlas

común, para que tuviesen las mismas dimensiones y los mismos vóxeles representasen la misma información anatómica.

Así, la primera fase de preprocesamiento de las imágenes RM se basaba en normalización espacial. En primer lugar, se lleva a cabo *resampling* para normalizar las imágenes a una resolución isotrópica de $2mm^3$ por vóxel. El procedimiento reduce las dimensiones de todas las imágenes. Posteriormente, se registran las imágenes a un atlas por medio de registrado rígido. El atlas utilizado ha sido el MNI 305 (Collins, Neelin, Peters, y Evans, 1994; Evans, 1992; Evans et al., 1993, 1992). Este está construido a partir de la media de 305 imágenes T1 ponderadas, y al normalizarlo espacialmente a $2mm^3$ por vóxel obtiene una resolución final de 78x110x86. Por tanto, las imágenes finales tendrán esta resolución.

El procedimiento de *resampling* se basa en una función tomada de códigos de ejemplo de la librería DLTK (Pawlowski et al., 2017). Por su parte, el procedimiento de registro a un atlas se ha llevado a cabo por medio de SimpleElastix (Marstal, Berendsen, Staring, & Klein, 2016), una librería para registro de imágenes médicas construida sobre SimpleITK (Beare, Lowekamp, & Yaniv, 2018; Lowekamp, Chen, Ibáñez, & Blezek, 2013; Yaniv, Lowekamp, Johnson, & Beare, 2018).

Por último, se ha llevado a cabo un proceso de extracción del cráneo para eliminar la información irrelevante de las imágenes y dejar únicamente el tejido cerebral. Para esto se ha hecho uso de FSL BET, un software de neuroimagen diseñado para este propósito (Smith, 2002). Su parámetro clave es el umbral de intensidad fraccionaria, que básicamente mide la agresividad del algoritmo a la hora de eliminar los componentes de la imagen que no pertenecen al cerebro. Un valor muy pequeño podría dejar demasiada información del cráneo, pero un valor muy grande podría hacer que se eliminara también parte del cerebro. En la base de datos utilizada, diferentes imágenes habían sido preprocesadas de diversas formas, por lo que no existe un valor de intensidad fraccionaria ideal para todo el *dataset*. Tras diversas pruebas, se tomó la decisión de utilizar un valor de 0.2, que era capaz de mantener un balance muy decente con la totalidad de las imágenes.

En este punto, cabe destacar que FSL BET es una herramienta que funciona sobre la línea de comandos de los sistemas operativos UNIX. Para poder integrarla en el código Python, se ha hecho uso de Nipype, una librería de código abierto diseñada para el tratamiento de neuroimagen con Python (Gorgolewski et al., 2011). Contiene una función (*interfaces.fsl.BET*) que interactúa directamente con la línea de comandos para ejecutar FSL BET.

En la Ilustración 8 se pueden observar tres planos o cortes axial, coronal y sagital de una imagen preprocesada. En este caso apenas hay rastro del cráneo y no se ha perdido tejido cerebral, pero en otros se puede dar una de estas situaciones debido a que el umbral de intensidad fraccionaria no funciona igual de bien con todas las imágenes.

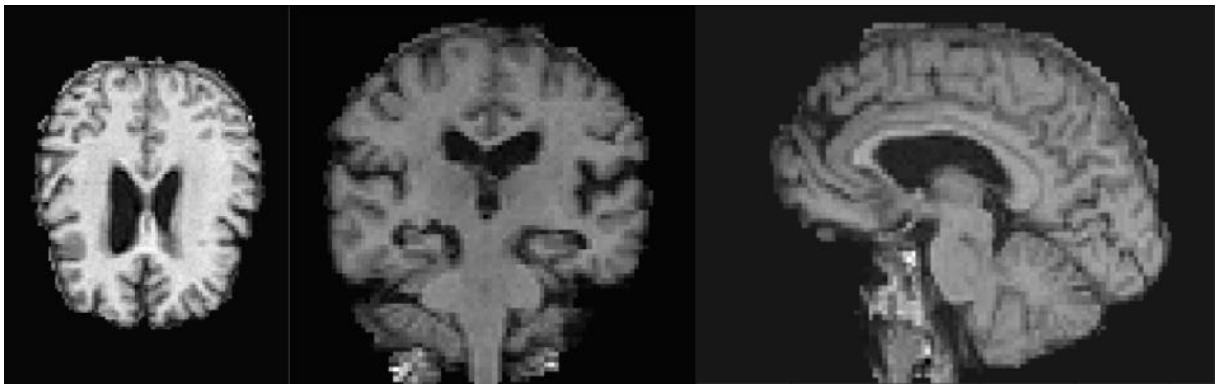

Ilustración 8. Planos axial, coronal y sagital, respectivamente, de una imagen registrada y con el cráneo extraído

Las técnicas de preprocesamiento utilizadas pueden ser bastante complejas, pero se llevan a cabo con un enfoque extremadamente simple. Se utilizan librerías y herramientas de código abierto y se aplican indistintamente a todas las imágenes, buscando obtener los mejores resultados posibles, pero sin tratar cada imagen por separado. Esta es una buena forma de probar hasta qué punto las herramientas disponibles hoy en día posibilitan el despliegue de sistemas de este tipo, ya que llevar a cabo el tratamiento de las imágenes de forma individual puede no ser asumible para científicos de datos sin experiencia profunda en temas de radiología.

TRANSFORMACIÓN A 2D

El primero de los modelos implementados ha sido una red Inception V3 pre-entrenada en Imagenet. Este tipo de redes están diseñadas para trabajar con imágenes RGB, las cuales tienen dos dimensiones para la estructura y una tercera para la información de color. Por tanto, no pueden trabajar con imágenes de 78x110x86 vóxeles.

Se ha implementado un algoritmo sencillo para la transformación de este tipo de imágenes a dos dimensiones, inspirado en el mismo procedimiento llevado a cabo por Ding et al. (2018), pero bastante más sencillo. Se trata de realizar múltiples cortes axiales (horizontales) y colocarlos sobre un mismo plano para construir una imagen bidimensional. Concretamente, se han tomado 16 cortes diferentes y se han colocado sobre una “matriz” de dimensiones 4x4. La imagen bidimensional resultante se replica tres veces para adoptar dimensiones RGB. Así, 16 cortes de tamaño 110x86 resultan en una imagen de 440x344x3.

Previamente a realizar los cortes, las imágenes 3D se normalizan para tener una media igual a cero y una desviación estándar igual a la unidad. Es decir, se les aplica un procedimiento de *whitening*. Posteriormente, se seleccionan los 16 cortes entre la altura axial de 28 y 60, que tras múltiples pruebas se comprobó que representaban suficiente información. Los cortes de una altura superior a 60 mostraban la parte superior del cerebro muy alejada, y una altura menor de 28 mostraba únicamente regiones del tronco encefálico. Los 16 cortes se distribuyen a distancias iguales entre sí, es decir, se extrae uno de cada dos planos.

En la Ilustración 9 se puede observar un ejemplo de una imagen convertida a dos dimensiones.

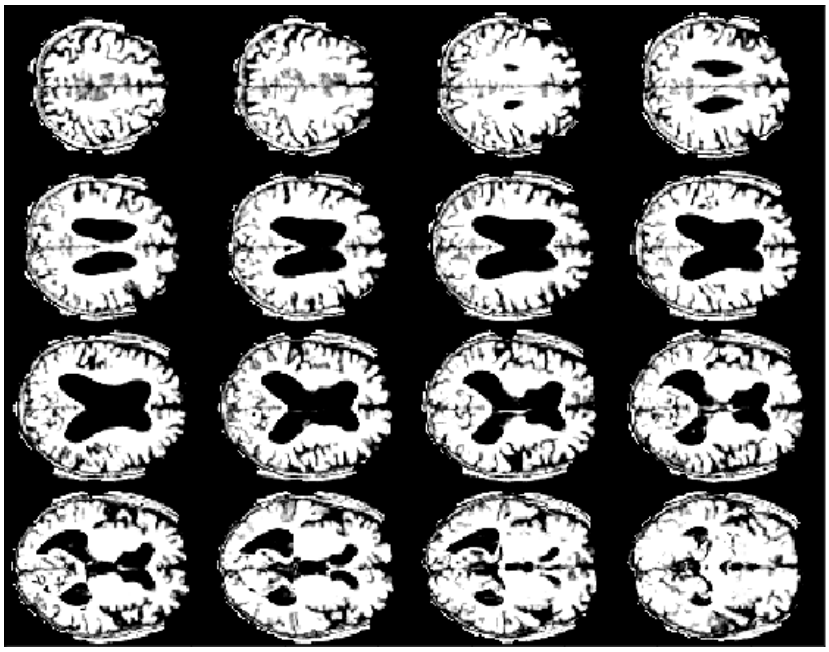

Ilustración 9. Imagen resultada de la conversión a dos dimensiones

3.1.2 TFRecords

Las imágenes en tres dimensiones son bastante más complejas que las imágenes RGB comunes. La complejidad a la hora de tratarlas es mucho mayor, y además ocupan mucho más espacio en memoria. Al utilizar más de 3000 imágenes, los 12GB de RAM disponibles en Google Colaboratory se quedan muy cortos, por lo que se debe buscar una forma alternativa de alimentar los modelos.

Una opción muy utilizada en Keras sería hacer uso de los generadores nativos de Python, leyendo la información desde el disco duro en lugar de desde la memoria RAM. Sin embargo, este procedimiento es mucho más lento y añade un cuello de botella incluso

mayor al que se tenía anteriormente. A la complejidad de tratar imágenes en tres dimensiones, se le estaría añadiendo la complicación de tener que cargarlas desde el disco duro. De hecho, en las pruebas realizadas con un generador de Keras personalizado, el entrenamiento era demasiado lento, llegando a necesitar de hasta cinco minutos por cada época (*epoch*). En entrenamientos de más de 50 épocas, esto es inasumible con los recursos disponibles.

Para estos casos, Tensorflow pone a disposición de los desarrolladores un formato de almacenamiento en disco duro basado en ficheros *tfrecords*. Los objetos se serializan y se almacenan en ficheros que son de lectura rápida para los tensores de Tensorflow. Esto requiere de duplicar la información en el disco duro para guardarla en este formato, pero también es mucho más rápido que hacer uso de los generadores nativos («Using TFRecords and tf.Example», s. f.). Este formato también es compatible con los modelos de Keras.

De esta forma, se han creado múltiples funciones en Python para organizar y almacenar la información en este formato. Posteriormente, estos ficheros se leen durante el entrenamiento y se pasa la información a los modelos *batch* a *batch*. Durante la lectura, se aplica *whitening* sobre las imágenes, por medio de operadores de tensores («tf.image.per_image_standardization», s. f.).

3.2 Modelos implementados

3.2.1 Reajuste de Inception V3 pre-entrenada

En primer lugar, siguiendo las tendencias exitosas de los últimos años de adaptar redes pre-entrenadas en Imagenet para aplicaciones médicas (Ding et al., 2018; Esteva et al., 2017; Gulshan et al., 2016; Vesal et al., 2017), se utilizaron las imágenes adaptadas a dos dimensiones para reajustar una red Inception V3 (Szegedy et al., 2015). El módulo *applications* de Keras permite descargar una red de este tipo con los pesos de Imagenet e ignorar la capa final. Esta se ha añadido posteriormente de forma manual con 512 unidades y una regularización mediante *dropout* del 80%. Finalmente, se añade una capa *softmax* para clasificación con tres unidades de salida (AD/MCI/CN).

En una primera fase de entrenamiento, se congelan todas las capas del modelo base y se entrenan únicamente las dos capas finales añadidas de forma manual. Esto se lleva a cabo durante 10 *epochs* con el optimizador Adam (Kingma y Ba, 2014) y una ratio de aprendizaje

(*learning rate*) de 10^{-4}. También se añade una constante de caída de esta ratio de 10^{-3}. Este valor es bastante agresivo, pero es útil al estar entrenando únicamente la última capa.

Posteriormente, se lleva a cabo *fine-tuning* de la totalidad de la red. En este caso, también se utiliza Adam con una ratio de aprendizaje de 10^{-4}, pero se disminuye la constante de decaída a 10^{-7}. Se mantiene el tamaño de batch de 8 elementos. El entrenamiento se alarga durante 70 *epochs*, lo cual es algo más de lo necesario, pero sirve para hacerse una idea de cómo evoluciona el modelo. Se utiliza la entropía cruzada categórica (*categorical cross-entropy*) como función de coste o pérdida y la exactitud (*accuracy*) como métrica de rendimiento auxiliar. En la Ilustración 10 se observa la evolución de ambos parámetros.

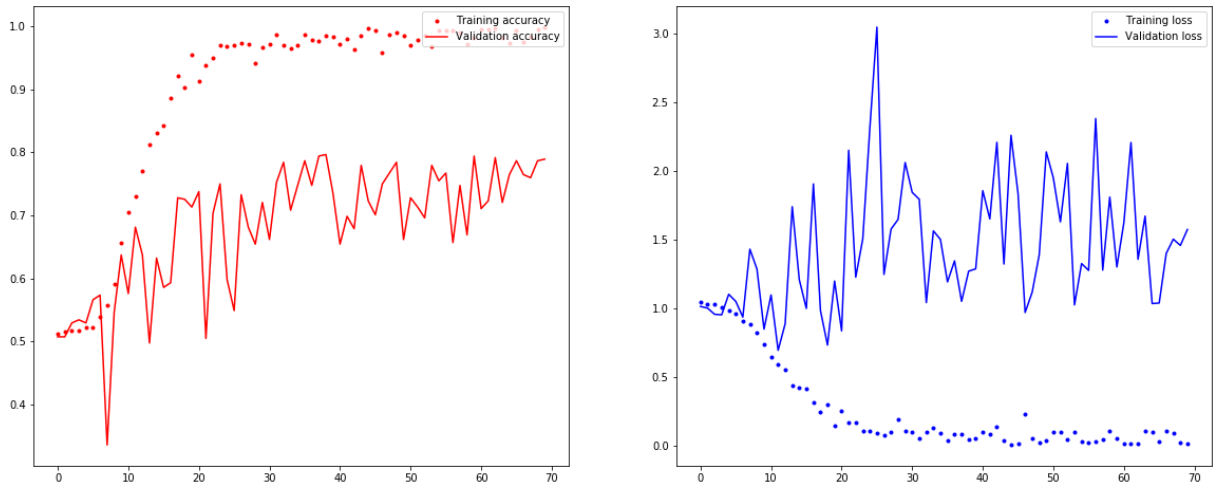

Ilustración 10. Exactitud (izquierda) y pérdida (derecha) para Inception V3 sin balanceo de clases. El entrenamiento son los puntos y la validación son las líneas continuas

Tanto la exactitud como la pérdida tienen unos valores bastante irregulares. Esto probablemente se debe a un factor clave: el desbalanceo de clases. Esto complica ligeramente la evaluación con el conjunto de validación. El entrenamiento, por su parte, evoluciona de forma algo más consistente. No obstante, balancear las clases apenas logra mejorar los resultados (véase la Ilustración 11). Conviene recordar que utilizar un conjunto de datos desbalanceado suele ser mejor idea que balancear los datos eliminando observaciones, es decir, que el *undersampling* no suele ser una buena estrategia (Mazurowski et al., 2008).

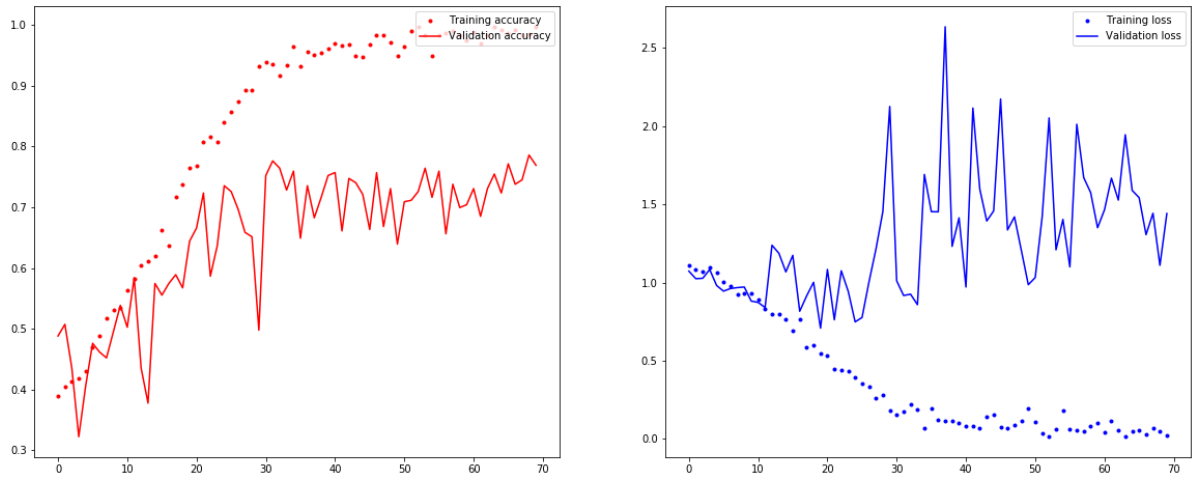

Ilustración 11. Exactitud (izquierda) y pérdida (derecha) para Inception V3 con balanceo de clases mediante *undersampling*. El entrenamiento son los puntos y la validación son las líneas continuas

La exactitud se utiliza únicamente como medida auxiliar. En investigación médica, no es una buena métrica para valorar el rendimiento de un modelo (Mazurowski et al., 2008). En su lugar, es muy común que se utilice el área bajo la curva ROC (AUC) para evaluar el rendimiento con cada una de las clases (Ding et al., 2018; Esteva et al., 2017; Gulshan et al., 2016; Hosseini-Asl, Gimel'farb, et al., 2016; Korolev et al., 2017; Vesal et al., 2017). Por este motivo, se ha diseñado un método para computar el área bajo la curva de cada una de las clases, dado un modelo entrenado y un conjunto de prueba. Este método se apoya en las métodos de evaluación presentes en la librería *scikit-learn* (Pedregosa et al., 2011).

La Ilustración 12 muestra el área bajo la curva para un modelo entrenado con las clases no balanceadas. La media pasa de 0.9, y obtiene resultados muy decentes con las clases CN y AD. Tiene más dificultades con la clase MCI, que no llega a 0.9.

Estos resultados sugieren que el modelo podría hacerlo bastante mejor si se reduce la clasificación a binaria (CN vs AD). De hecho, Korolev et al. (2017) construyen directamente múltiples modelos de clasificación binaria. No obstante, conviene destacar que utilizan modelos 3D, más complejos, y aún así el reajuste de la Inception V3 multiclase realizado en este trabajo logra mejores resultados. También es cierto que toman un enfoque diferente, utilizando un conjunto de datos mucho menor, pero su premisa de simplicidad también se persigue en este trabajo.

En la Ilustración 13 se muestra el área bajo la curva para un modelo entrenado con clases balanceadas mediante *undersampling*. Los resultados son inferiores, aunque no por un margen demasiado amplio. De nuevo, esta técnica no suele mejorar los resultados, ni siquiera con conjuntos de datos muy desbalanceados.

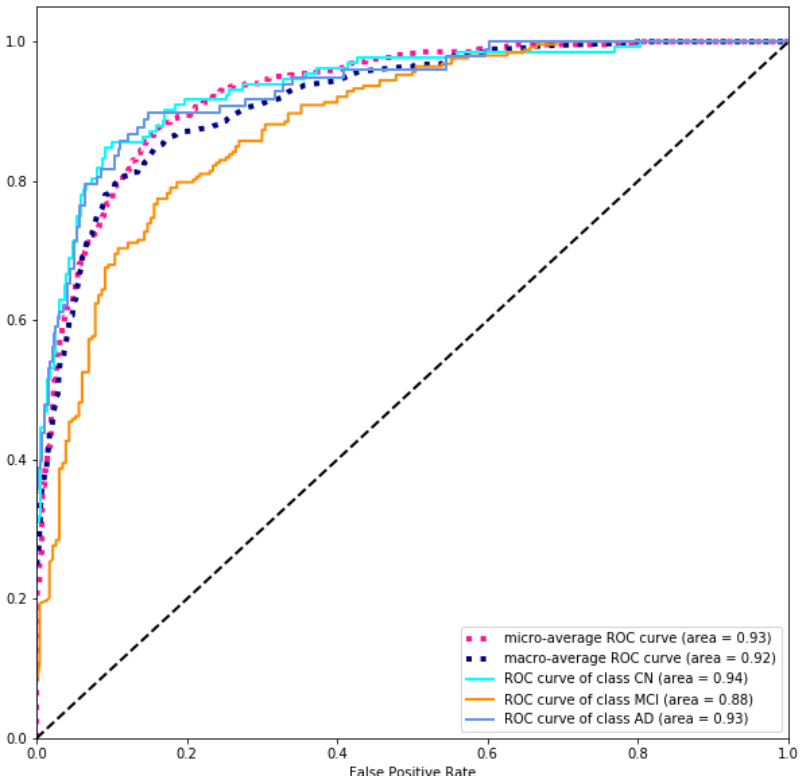

Ilustración 12. AUC para Inception V3 sin balanceo de clases

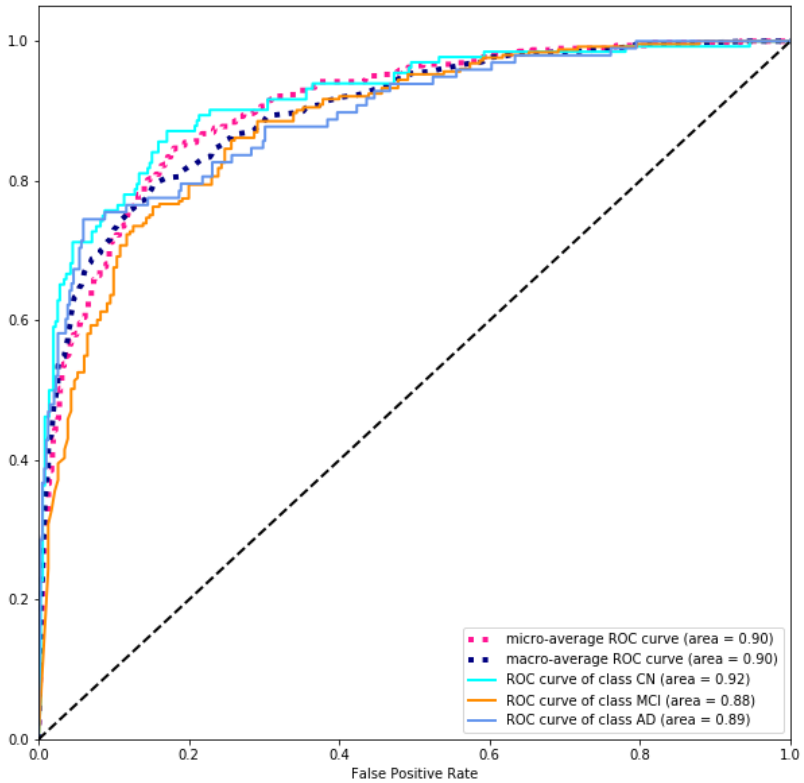

Ilustración 13. AUC para Inception V3 con balanceo de clases

3.2.2 Redes neuronales convolucionales 3D

El rendimiento de la Inception V3 se compara con el obtenido construyendo una red convolucional desde cero, y que trabaja con las imágenes directamente en tres dimensiones. Inmediatamente, plantear la construcción de un modelo semejante sugiere dos cosas: que el entrenamiento será notablemente más lento al utilizar la información 3D, y que la cantidad de imágenes disponibles puede quedarse corta. En general, cuanto más compleja es la red y más información tiene que procesar, mayor es el número de observaciones que necesita para aprender.

La arquitectura seleccionada ha sido la de una red residual (ResNet), pero adaptada para trabajar con imágenes en tres dimensiones. Este tipo de arquitectura, bautizada VoxResNet, fue propuesta por H. Chen et al. (2016). En la librería DLTK, se ofrece un modelo de este tipo para acometer problemas de clasificación y regresión («dltk.networks.regression_classification package», 2017). Puesto que en este trabajo se utiliza Keras por simplicidad, se ha recurrido a una implementación disponible en un repositorio público de GitHub, que implementa una clase para la construcción sencilla de este tipo de modelos (JihongJu, 2017).

Esta implementación permite seleccionar la profundidad que se desea que tenga la ResNet 3D. Dispone de métodos para construir redes de profundidades 18, 34, 50, 101 y 152. Inicialmente, se ha hecho uso de la red más pequeña, con tan solo 18 capas residuales, debido a la complejidad del entrenamiento de modelos con imágenes en 3D. Se ha buscado mantener los modelos lo más simples posible, pues el entrenamiento es más rápido y es más sencillo evitar el *overfitting*.

Al modelo se le ha añadido una capa final de 512 unidades y *dropout* del 80%, y una capa *softmax* adicional para clasificación, de forma similar a como se hacía con la Inception V3. El entrenamiento se ha ejecutado durante 50 *epochs* con el optimizador Adam, una ratio de aprendizaje de 10^{-5} y un *batch* de tamaño 8. La función de pérdida ha sido la *categorical cross-entropy*, con el *accuracy* como métrica auxiliar. El modelo cae en *overfitting* con relativa facilidad, por lo que se estuvieron realizando experimentos con el factor de regularización. Se trata del valor de regularización L2 utilizado en las capas residuales.

Los resultados obtenidos por esta red guardaron ciertas semejanzas con los obtenidos por la Inception V3. Principalmente en la irregularidad de los resultados obtenidos con el conjunto de validación. En la Ilustración 14 se puede observar la evolución de la exactitud y la pérdida, tanto en validación como en entrenamiento, al utilizar un factor de regularización de 0.05. La curva de aprendizaje no es nada suave y da saltos muy notables, dificultando la

toma de decisiones a la hora de ajustar los hiperparámetros. Aumentar la regularización no sólo no mejoraba esta situación, sino que además terminaba otorgando peores resultados con el conjunto de prueba.

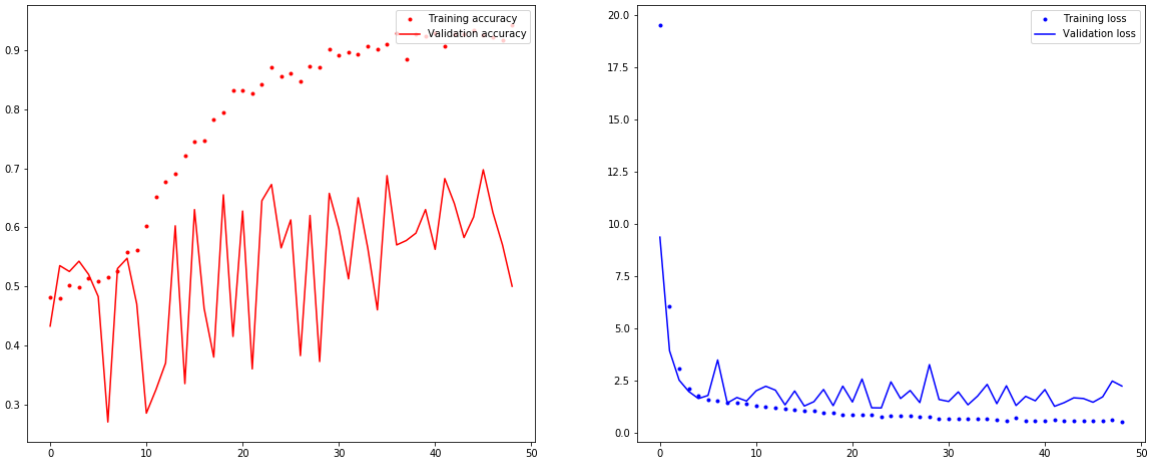

Ilustración 14. Evolución de la pérdida y la exactitud de una ResNet3D-18 con un alto factor de regularización (0.05)

Reajustando ligeramente este factor de regularización, se acabó generando un modelo de 18 capas que obtenía el rendimiento con el conjunto de prueba que se puede observar en la Ilustración 15. El AUC está en torno a cinco puntos por debajo del rendimiento obtenido con el reajuste de la Inception V3. De nuevo, tiene más problemas con la clase MCI que con el resto. En pruebas posteriores, no se logró superar este rendimiento.

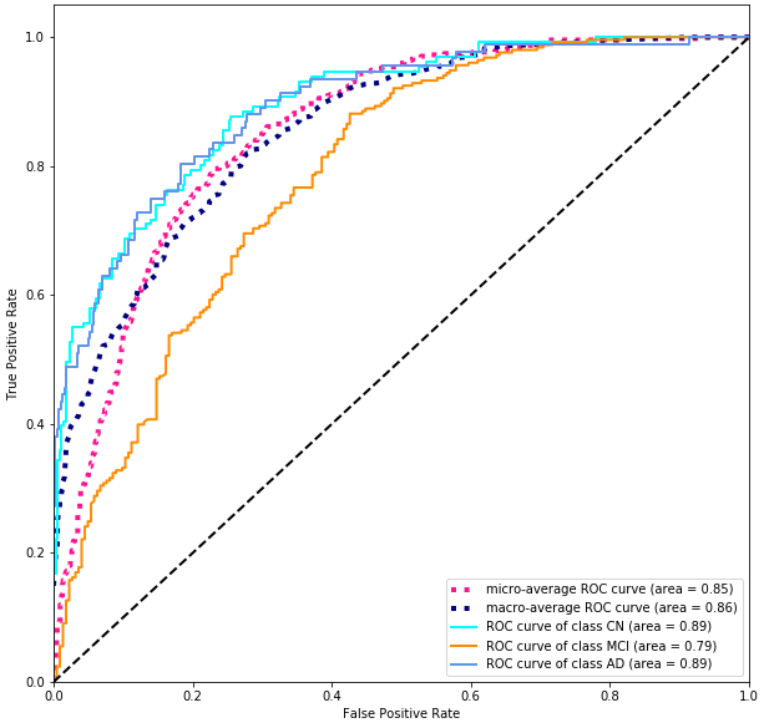

Ilustración 15. AUC con una ResNet3D-18 y un factor de regularización de 0.03

Posteriormente, se buscó mejorar el rendimiento de la red aumentando su profundidad. Eso sí, esto se hizo de forma completamente exploratoria, pues el entorno de desarrollo utilizado no era idóneo para realizar múltiples pruebas. El objetivo era hacerse una idea de hasta qué punto la red podría mejorar incrementando su profundidad, y manteniendo el conjunto de datos.

Una ResNet 3D de 50 capas fue totalmente inasumible. Para completar las tres primeras épocas del entrenamiento, se necesitaron cerca de 30 minutos. Al tratarse de una red más del doble de grande, el entrenamiento debía desarrollarse de forma incluso más lenta y cuidadosa que con la red de profundidad 18. Un entrenamiento de sólo 10 épocas más (60) habría requerido de 10 horas.

Por otro lado, el entrenamiento de una red de profundidad 34 era ligeramente más factible. De nuevo, no buscando maximizar el rendimiento por encima de todo, sino únicamente explorando su potencial. Para ello, se entrenó en las condiciones similares a las de la red de 18 capas, es decir, añadiendo dos capas finales de 512 y 3 unidades respectivamente, y con la misma ratio de aprendizaje y optimizador. Únicamente se incrementó el factor de regularización a 0.04 y se alargó el entrenamiento hasta 70 *epochs*. Los resultados se pueden observar en la Ilustración 16. A partir de los 40 *epochs*, el modelo cae rápidamente en *overfitting* (Ilustración 17).

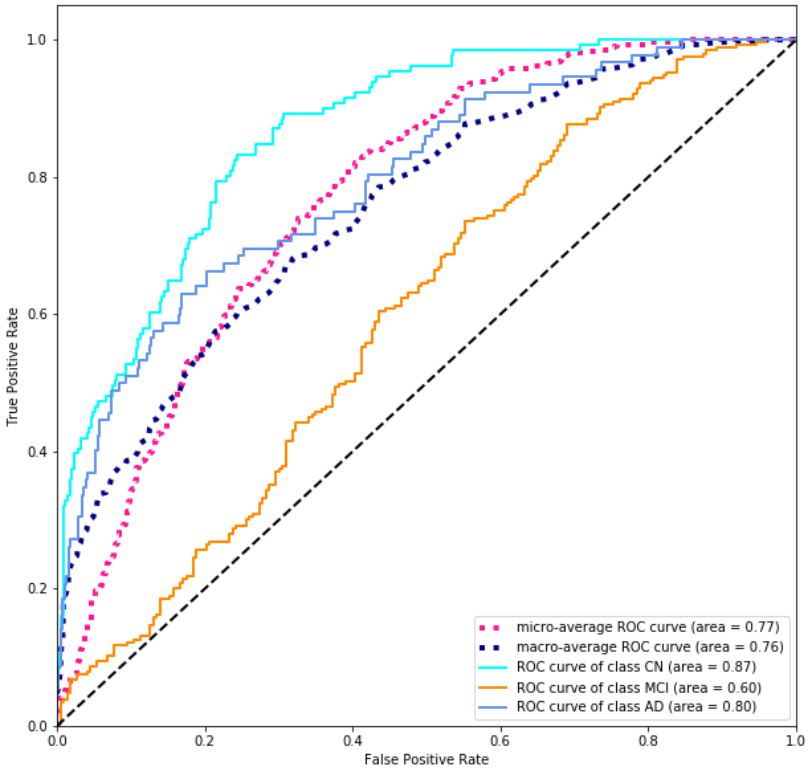

Ilustración 16. AUC para ResNet3D-34 con factor de regularización de 0.04

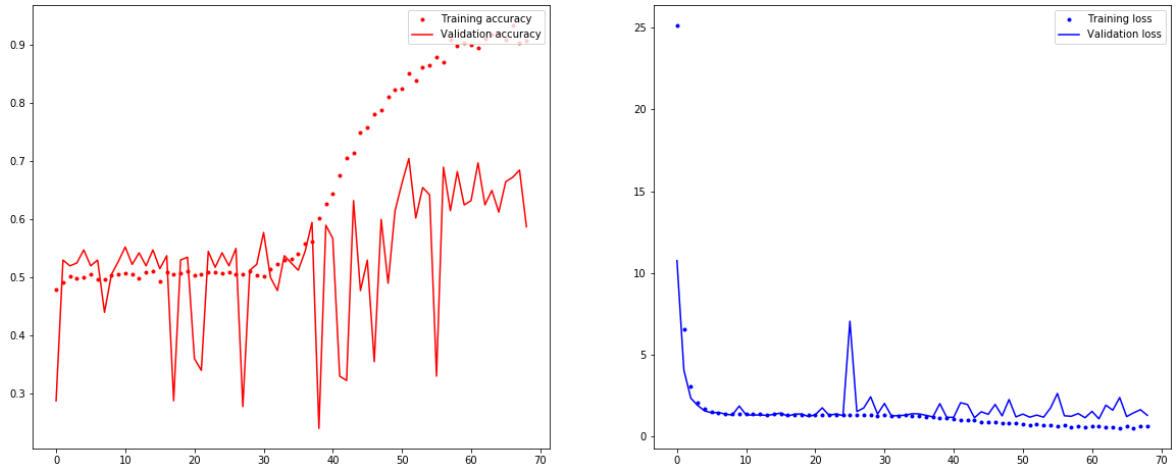

Ilustración 17. Entrenamiento por época para una ResNet3D-34 con factor de regularización 0.04

El rendimiento no es especialmente destacable ni prometedor. Es igual de irregular que con los modelos anteriores, lo cual era de esperar, pero tampoco es capaz de obtener valores de exactitud más altos ni valores de pérdida más bajos. Utilizando criterios de parada (*early stopping*) basados en los valores de las métricas se podría extraer un modelo relativamente potente, pero no hay indicios que sugieran que el modelo de 34 capas es superior al de 18 capas, pese a que es considerablemente más complejo. Probablemente haría falta un conjunto de datos mayor para obtener mejoras notables.

En la Ilustración 18 se incluyen los resultados obtenidos al disminuir ligeramente el factor de regularización. Son notablemente mejores, pero lejos de los obtenidos con la Inception V3 o incluso con la mejor ResNet3D de 18 capas.

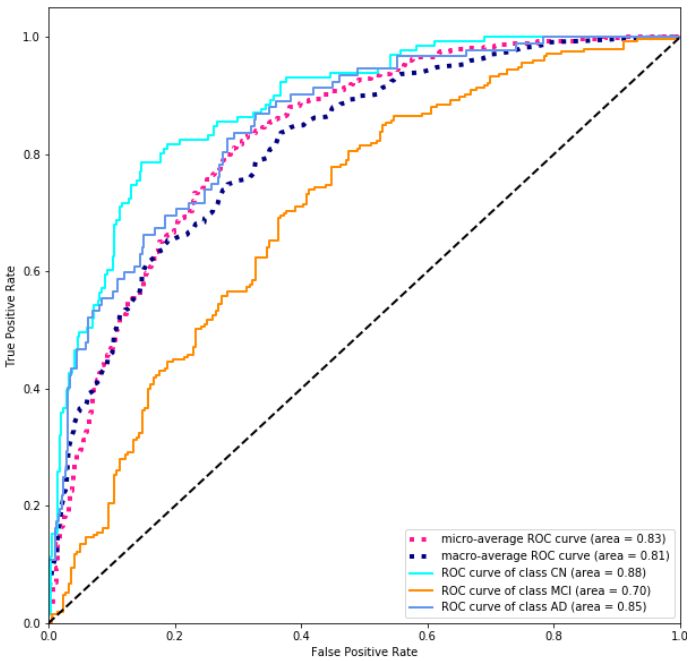

Ilustración 18. AUC para ResNet3D-34 con factor de regularización de 0.03

3.3 Análisis

Los resultados obtenidos han sido, a grandes rasgos, diferentes de lo que se podría haber esperado en un inicio. Aunque es cierto que el trabajo con las imágenes en 3D vaticinaba una complejidad mucho mayor y la necesidad de más datos para entrenamiento, el uso de la información completa de las imágenes sugería que los resultados serían mejores. Sin embargo, los modelos en 2D derivados del *fine-tuning* de una red Inception V3 han sido superiores a los que se han podido obtener con cualquiera de las ResNet 3D. Esto es aún más sorprendente teniendo en cuenta que el algoritmo de conversión a 2D no hace uso de ningún tipo de criterio clínico o científico, sino que simplemente extrae planos horizontales situados a la misma distancia los unos de los otros, dentro de un rango estimado de forma manual.

Ding et al. (2018) logran algunos de los resultados más avanzados de la literatura siguiendo un procedimiento similar, pero apoyándose en criterios más sofisticados para la extracción de planos. Como resultado, el modelo se encuentra alrededor de cinco puntos por encima del mejor obtenido en este estudio en la detección de la clase AD con el conjunto de prueba, obteniendo un AUC de 0.98 frente a 0.93 (consultar Ilustración 12). También encuentran las mayores dificultades a la hora de clasificar la clase MCI, lo cual ha sido una constante en este trabajo. Teniendo en cuenta que el conjunto de datos utilizado en ese trabajo está considerablemente mejor balanceado, con tan sólo 100 ejemplares más de MCI frente a CN, se podría contemplar la posibilidad de que estas dificultades sean inherentes a la propia naturaleza de las clases.

No obstante, existe una diferencia clave entre dicho estudio y el realizado por el autor: el tipo de imágenes utilizadas. Ding et al. (2018) utilizan imágenes 18FGD-PET, obtenidas mediante un tipo de tomografía por emisión de positrones y que poseen una dimensión temporal. Es decir, que se trata de imágenes en cuatro dimensiones, siendo la cuarta la relativa al tiempo. Para extraer un único volumen 3D, hicieron uso de una técnica conocida como análisis de componentes conectadas, “seleccionando las secciones más craneal y más caudal que representan más de $100x100mm^2$ de parénquima cerebral” (Ding et al., 2018, p. 2). Por otro lado, en este estudio se ha hecho uso de imágenes de resonancia magnética de la modalidad T1 ponderada.

Esta diferencia es fundamental por dos motivos principales. El primero es la evidente simplicidad del enfoque de este trabajo, donde se extraen los planos por medio de un procedimiento sencillo y accesible a cualquier científico de datos. El segundo es la diferencia de coste en la obtención de imágenes RM frente a las TEP. La tomografía por emisión de

positrones es capaz de capturar la forma en la que funciona cerebro, lo que sería la actividad cerebral, pero es más costosa y utiliza materiales radioactivos que pueden ser muy nocivos para el paciente. Por el contrario, la resonancia magnética puede capturar únicamente la estructura del cerebro, no su funcionamiento, pero es más barata e inocua para el sujeto (Feldman, 2004; Hoffman et al., 2000; Nordberg et al., 2010; Pichler, Kolb, Nagele, y Schlemmer, 2010). Por tanto, se puede afirmar que sería más deseable obtener un modelo capaz de diagnosticar el Alzheimer a partir de las IRM, pues son más baratas de obtener y menos perjudiciales para el paciente.

Aunque es cierto que el rendimiento obtenido reajustando la Inception V3 queda algo lejos del demostrado por Ding et al. (2018) con imágenes TEP, es importante tener en cuenta el enfoque basado en la sencillez mantenido en este trabajo. Los procedimientos de registro de imágenes y extracción del cráneo llevados a cabo para preprocesar las IRM no han sido extremadamente cuidadosos, sino que se ha hecho uso directo de herramientas de código abierto. Especialmente en la extracción del cráneo podrían obtenerse mejoras considerables al utilizar un umbral de intensidad fraccionaria ajustado a cada una de las imágenes. Teniendo esto en cuenta, utilizar imágenes de resonancia magnética se convierte en una alternativa muy viable y que merece ser explorada en mayor profundidad, sobre todo según más datos se vayan haciendo disponibles con el tiempo.

En la bibliografía disponible actualmente, las investigaciones se han centrado en maximizar el rendimiento a toda costa, aunque eso suponga recurrir a procedimientos complejos de preprocesamiento de la información o a arquitecturas diseñadas manualmente. La única excepción a esta regla es el trabajo de Korolev et al. (2017), que también prioriza la simplicidad y la aplicación del *Deep Learning* de forma más directa. Por tanto, tiene sentido comparar los resultados obtenidos con los presentados en esa publicación. Previamente, es importante destacar que utilizan exclusivamente redes 3D y clasificación binaria, construyendo múltiples modelos para enfrentar cada una de las clases dos a dos. Tan sólo el modelo que predice AD vs NC obtiene unos resultados comparables a los obtenidos en este trabajo por la Inception V3, con un AUC de 0.88 y 0.87 (de su VoxCNN y ResNet3D, respectivamente) frente a las medidas presentadas en la Ilustración 12. Se puede observar un resumen de los AUC medios obtenidos en dicho trabajo en la Tabla 5.

Tabla 5.
AUC medias obtenidas en Korolev et al. (2017) (Korolev et al., 2017)

	AUC VoxCNN	AUC ResNet
AD vs NC	0.88	0.87
AD vs EMCI	0.66	0.67
AD vs LMCI	0.61	0.62
LMCI vs NC	0.67	0.65
LMCI vs EMCI	0.47	0.52
EMCI vs NC	0.57	0.58

En este caso, dividen la clase MCI en dos clases diferentes, distinguiendo entre MCI temprano o *early* MCI (EMCI) y MCI tardío o *late* MCI (LMCI). La idea, sin embargo, es la misma, y se pueden extraer conclusiones similares en cuanto a que MCI es la clase más complicada de manejar

Los resultados obtenidos por su ResNet son ligeramente inferiores a algunos de los obtenidos con la ResNet 3D presentada en este trabajo, aunque hay que tener en cuenta que se están manejando valores medios. Si se utilizan los valores máximos, la clasificación AD vs NC ganaría 8 y 7 puntos en la VoxCNN y la ResNet, respectivamente. En cualquier caso, se partirá de la premisa de que los resultados de ambos estudios son comparables. Esto tiene una serie de implicaciones interesantes.

En primer lugar, no es necesario crear múltiples modelos para llevar a cabo clasificaciones binarias, sino que la clasificación multiclase es una vía de desarrollo igual de factible. La única razón clara para hacer uso de clasificación binaria sería el poder tratar más cuidadosamente la clase MCI, que es con la que se tienen siempre mayores dificultades. Sin embargo, como puede observarse en la Tabla 5, son precisamente los modelos que tratan esta clase los que peores resultados terminan obteniendo, por lo que se debería buscar una mejor gestión de esta clase por otras vías.

En segundo lugar, destacar que las 231 imágenes utilizadas por Korolev et al. (2017) bastan para crear modelos comparables a los obtenidos en este trabajo, donde se utilizan más de 3000 imágenes. La diferencia radica en que esas 231 habían sido normalizadas y se les había extraído el cráneo manualmente, mientras que las 3000 han sido procesadas de forma más automática. Por tanto, sería muy complicado afirmar que las herramientas de registro y extracción del cráneo actuales son lo bastante avanzadas como para aplicarlas indistintamente a múltiples imágenes con ligeras variaciones entre ellas.

Finalmente, conviene destacar que este tipo de modelos tienen aún mucho camino por recorrer para poder ser desplegados en entornos clínicos reales. Observando los resultados de forma directa, se podría decir que todavía tienen un amplio margen de error, y por lo

tanto su diagnóstico no debería tenerse en cuenta. Por otro lado, también es cierto que otros estudios han llevado a cabo una comparación directa con radiólogos y han emitido conclusiones sorprendentes, afirmando que los modelos son significativamente superiores (Ding et al., 2018). Por tanto, se podría considerar el hecho de que no necesariamente se busca la perfección en estos modelos, sino simplemente el superar la capacidad de los radiólogos. En tal caso, los mejores modelos deberían compararse siempre de forma directa con profesionales humanos para extraer una medida de calidad real, y estimar si realmente es factible su despliegue en entornos reales.

3.4 Recomendaciones

Tras el análisis completo de los resultados y la comparación con otros trabajos importantes de los últimos años, se pueden realizar una serie de recomendaciones a un científico de datos sin conocimientos profundos de medicina que busque construir un sistema para diagnóstico CADx del Alzheimer. Este era el tercer y último objetivo específico de este trabajo.

Antes que nada, habría que señalar que el uso de redes neuronales convolucionales es la opción inmediata. En el capítulo 2 se hizo mucho hincapié en la superioridad de estos modelos frente a otras opciones como las SVM. De hecho, las CNN acaparan la práctica totalidad de las aplicaciones en análisis de imágenes médicas en general, no sólo en el diagnóstico del Alzheimer.

Hablando del material a utilizar, se podría decir que la base de datos del ADNI contiene imágenes suficientes y de calidad como para construir buenos modelos. Esto es algo que se podía concluir a partir de trabajos anteriores (Korolev et al., 2017). Ahora bien, tratar esta información es algo más complicado, pues las herramientas de código abierto disponibles no son perfectas, sobre todo a la hora de realizar la extracción del cráneo. Lo recomendable sería tratar cada grupo de imágenes de forma individual, buscando adaptar el umbral de intensidad fraccionaria (en el caso de FSL BET) según convenga. El problema es que la única forma de gestionar si el valor de umbral es el correcto es observando directamente el resultado de la extracción, y esto puede ser muy costoso para conjuntos de imágenes muy grandes. Por otro lado, las herramientas de registro de imágenes funcionan bastante bien. Son suficientes para tratar las imágenes de forma automática, siempre y cuando se registren a un atlas de la misma modalidad (T1 ponderado, en el caso de este trabajo) y se normalicen todas las imágenes a la misma resolución espacial ($2mm^3$ en este trabajo).

En este punto, conviene destacar que, a mayor resolución espacial, las imágenes resultantes serán más grandes y cada vóxel representará regiones más pequeñas. Esto hace suponer que utilizar una resolución isotrópica de $1mm^3$ sería mejor, pero el mayor tamaño de las imágenes haría el entrenamiento más lento. Por tanto, es una cuestión de tiempo y capacidad de procesamiento.

Con respecto a la estructura de las imágenes, sería mucho más recomendable transformar la información a dos dimensiones, aunque fuera con un procedimiento muy sencillo como el que se ha utilizado en este trabajo. Las redes en 3D son mucho más complicadas de entrenar, son más lentas y consiguen peores resultados. Las redes 2D, en cambio, entrenan más rápidamente, y existe una gran variedad de modelos avanzados sobre los que hacer *fine-tuning*. Además, son mucho más intuitivas, en el sentido en que requieren orientar el problema de forma similar a como lo haría un radiólogo, examinando múltiples planos de forma individual. Este último punto es muy importante, pues es más sencillo atacar un problema con *Deep Learning* tomando como inspiración la forma en la que un ser humano lo soluciona.

En cuanto a la modalidad de las imágenes, tanto IRM como TEP se pueden utilizar. Las segundas no requieren de extracción de cráneo y han demostrado resultados excelentes en otros trabajos, pero son más caras de obtener y más perjudiciales para el paciente. En cambio, las IRM son baratas e inocuas para los pacientes, pero requieren de un preprocesamiento más cuidadoso. Al final, la elección se reduce a la disponibilidad y a las preferencias de los interesados.

A continuación, se aborda la forma en la que se debería llevar a cabo el entrenamiento de la red. En general, utilizar una arquitectura reconocida, como Inception o ResNet, debería ser la opción por defecto, ya que han demostrado ser muy buenas en una amplia variedad de aplicaciones. En el caso de este trabajo, se han logrado mejores resultados utilizando tamaños de *batch* y ratios de aprendizaje pequeños, alargando el entrenamiento en un gran número de *epochs*. Un optimizador adaptativo, como Adam, también sería muy adecuado. Además, los trabajos de los últimos años también han seguido estas directrices (Ding et al., 2018; Korolev et al., 2017).

También se plantea la cuestión de si llevar a cabo algún tipo de pre-entrenamiento por medio de autoencoders convolucionales. Esto ya se introdujo en el capítulo 2 como una opción muy utilizada en la literatura. Sin embargo, esta tendencia se ha ido perdiendo con el tiempo, especialmente debido al auge de los optimizadores adaptativos (Hellström, 2018). Además, utilizar este método obligaría a diseñar la red de forma manual, perdiendo la opción de ajustar una arquitectura potente como la Inception. Por lo tanto, la única razón de

peso para realizar pre-entrenamiento sería si se dispusiese de un conjunto de datos no etiquetado muy grande, del cual pudiesen extraerse características útiles, para luego hacer *fine-tuning* con los datos etiquetados. Comparar los resultados de un modelo semejante con los obtenidos en este trabajo sería una futura línea de investigación interesante.

Por último, es importante destacar que lograr resultados suficientemente buenos es extremadamente complicado. El tratamiento de la clase MCI tiende a perjudicar considerablemente el rendimiento de los modelos. Por esta razón, sería recomendable reducir la clasificación a binaria (AD vs NC) en aplicaciones críticas en las que haya que maximizar el rendimiento a toda costa.

Por encima de todo lo explicado hasta este punto, la recomendación principal sería el no desplegar un modelo de este tipo. El diagnóstico de una enfermedad es un tema extremadamente sensible, para el que no sólo son necesarios unos resultados excelentes, los cuales no se han alcanzado hasta el momento de la redacción de estas líneas, sino también un examen profundo de la fiabilidad de los modelos. En los próximos años, la creciente cantidad de datos disponibles debería aliviar estos problemas, pero por el momento se requiere de más y mejores investigaciones.

4 Conclusiones y líneas futuras

A continuación, se reúnen las conclusiones alcanzadas del trabajo realizado. A partir de los diferentes objetivos específicos prestablecidos en el capítulo 1, se resumen sus resultados. Respectivamente, estos objetivos han sido abordados en sucesivas secciones de este documento.

El primero de ellos consistía en encontrar los modelos más importantes del estado del arte. Esta tarea se aborda de forma detallada en el capítulo 2, concluyendo que las redes neuronales convolucionales son la opción inmediata en este aspecto, habiendo tan solo variaciones en la forma de entrenarlas. En este aspecto, tres métodos predominan en la bibliografía:

- Entrenamiento de redes neuronales de forma directa, utilizando las imágenes en tres dimensiones y arquitecturas propias e inicialización de pesos con métodos clásicos, sin pre-entrenamiento.
- Entrenamiento directo, con imágenes en tres dimensiones y arquitecturas propias e inicialización de pesos con métodos clásicos, pero utilizando autoencoders para pre-entrenar las redes. Esta opción se descarta para actividades posteriores debido a su pérdida de popularidad y a que no suele demostrar mejores resultados que la opción anterior.
- Reajuste de redes pre-entrenadas, que obliga a preprocesar las imágenes para convertirlas a dos dimensiones. Se utilizan arquitecturas que han logrado buenos resultados en otras tareas, como la competición ILVSRC. Es la opción por defecto en los últimos años.

De esta forma, la primera y la tercera técnica deberían ser las opciones por defecto para un científico de datos a la hora de implementar modelos de diagnóstico del Alzheimer (aunque también para diagnóstico de muchas otras enfermedades). Además, en principio son opciones relativamente sencillas de implementar y que no difieren mucho de los procedimientos seguidos en otros campos de aplicación.

Ahora bien, conseguir buenos resultados es una tarea bastante complicada debido a la naturaleza de las imágenes médicas. Por ello, en el segundo objetivo específico se buscaba implementar los modelos seleccionados en un escenario experimental. Todo ello con el fin de comprobar cómo de factible es su implementación por parte de profesionales sin conocimientos profundos de medicina, simplemente utilizando las herramientas disponibles.

Este objetivo se aborda en el capítulo 3. Aunque el análisis de los resultados corresponde al objetivo específico posterior, en este punto pueden extraerse algunas conclusiones menores. La implementación haciendo uso de las herramientas disponibles es, sin duda, factible y da lugar a la posibilidad de crear buenos modelos, sin necesidad de poseer conocimientos profundos de la enfermedad del Alzheimer. Se pueden seguir procedimientos comunes a otros campos de aplicación, tratando las imágenes de forma tradicional y ajustando hiperparámetros para buscar los mejores resultados posibles. Ese tratamiento tradicional se refiere a procedimientos de normalización de la intensidad que se utilizan en otras aplicaciones. Ahora bien, sí es fundamental familiarizarse, al menos, con herramientas de registro de imágenes médicas. Estas podrían ser desconocidas para muchos profesionales, pero su uso no requiere de conocimientos demasiado detallados de imágenes médicas. Algo similar ocurre con la extracción del cráneo.

El tercer objetivo específico consistía en analizar los resultados obtenidos en el escenario experimental, con el fin de emitir las recomendaciones oportunas a profesionales que busquen construir modelos de diagnóstico para el Alzheimer. En el capítulo 3, se incluyen dos secciones dedicadas a esto (3.33.4). A grandes rasgos, se concluye que la implementación de modelos bidimensionales, reajustando arquitecturas potentes como Inception V3, es la mejor opción. En las secciones referenciadas se lleva a cabo un análisis mucho más profundo de por qué, y de diversos detalles que se deben tener en cuenta a la hora de preprocesar las imágenes y entrenar estas redes.

En todo momento, este trabajo ha estado guiado por dos preguntas de investigación. Por un lado, ¿qué modelos, de entre las opciones posibles, logra mejores resultados? Y por otro, ¿cómo de factible es su implementación por parte de profesionales sin conocimientos profundos de medicina? La primera de estas preguntas se resuelve durante la fase exploratoria y se ha resumido también en estas conclusiones. No obstante, la segunda es probablemente las más interesante, y su respuesta se aborda brevemente en el apartado de las recomendaciones (3.4).

Teniendo en cuenta la sensibilidad de este campo de aplicación, y de acuerdo con los experimentos realizados en este trabajo, no se puede afirmar que este tipo de modelos se puedan llevar a producción de manera sencilla. Los resultados son bastante irregulares y, aunque pueden llegar a ser muy notables, en aplicaciones médicas probablemente se necesiten rendimientos que rocen la perfección. Es decir, valores de AUC que superen consistentemente el 0.99, o valores de *accuracy* que lleguen al 100%. Según se van haciendo disponibles más imágenes etiquetadas, se podría llegar a este nivel. Y aun

logrando estos resultados, sería necesario un detallado examen que asegure la no presencia de sesgos ni ningún tipo de errores ocultos en el modelo.

No obstante, el rendimiento de este tipo de modelos podría comprobarse de otra manera. Partiendo de la idea de que el uso de Inteligencia Artificial en medicina busca reducir el error humano, se podría valorar cómo de bueno es un modelo a partir de una comparación directa con radiólogos. En este escenario, un modelo podría no necesitar resultados prácticamente perfectos, sino simplemente igualar o superar a los radiólogos. Además, su uso se plantearía como un complemento al trabajo de los radiólogos, y no como un sustituto, permitiendo al profesional tener siempre la última palabra, pero permitiéndole reducir sus errores de forma notable. Así, la implementación de estos modelos por parte de profesionales sin conocimiento profundo de la enfermedad podría ser más razonable, aunque siempre sería necesario un procedimiento de evaluación basado en la comparación directa con radiólogos.

Finalmente, las contribuciones de este trabajo podrían resumirse en los siguientes puntos:

- Se analiza y resume de forma detallada el estado del arte en el análisis de imágenes médicas y el diagnóstico automático del Alzheimer, hasta principios del año 2019. Esto es una buena base para investigadores que busquen iniciarse en este campo.
- Se estima el rendimiento que los principales modelos de aprendizaje profundo son capaces de proporcionar en la tarea del diagnóstico del Alzheimer, haciendo uso de las herramientas disponibles hoy en día.
- Se proporcionan una serie de recomendaciones para profesionales que busquen implementar modelos de este tipo sin tener conocimientos profundos de medicina.
- Se concluye que sería muy complicado llevar a producción un modelo de este tipo de forma sencilla, al menos con los datos de los que se dispone hoy en día.

4.1 Líneas futuras

El análisis de imágenes médicas por medio de Inteligencia Artificial, así como el diagnóstico del Alzheimer, son campos muy extensos y complejos donde hay mucho espacio para investigación. En concreto, el trabajo realizado en este TFM plantea una serie de preguntas que merecerían ser abordadas en futuros trabajos.

Una primera línea ya se adelanta en el apartado 3.4. Se trata de reunir imágenes RM no etiquetadas para formar un conjunto de datos muy grande con el que llevar a cabo pre-

entrenamiento de una red convolucional, utilizando autoencoders convolucionales. Sería muy interesante comparar los resultados obtenidos por una red semejante con los obtenidos en este trabajo.

También sería interesante llevar a cabo una extracción del cráneo más cuidadosa en las imágenes RM. Concretamente, se podría ajustar el umbral de intensidad fraccionaria en base al procedimiento de normalización que se haya llevado a cabo sobre las imágenes previamente. En la base de datos del ADNI, las imágenes aparecen marcadas con una descripción de su modalidad, y de las posibles transformaciones aplicadas sobre ellas. Se trata de etiquetas como *B1-corrected* o *N3-scaled*. Estaría al alcance de cualquier científico de datos el agrupar las imágenes por medio de estas etiquetas, y tratar de encontrar el mejor umbral de intensidad fraccionaria para cada uno. Esto podría mejorar notablemente los resultados al extraer el cráneo, proporcionando información más limpia, detallada y útil a las redes convolucionales.

Otra posible extensión del trabajo realizado sería hacer uso de técnicas tradicionales de aumento de datos para incrementar el tamaño del conjunto de entrenamiento (*data augmentation*). Estas también están al alcance de cualquier profesional experimentado en la construcción de modelos de aprendizaje profundo. Esto, en principio, entraría en conflicto con el propósito de registrar las imágenes, haciendo que los píxeles o vóxeles de dos imágenes diferentes dejen de representar la misma información anatómica. Sin embargo, podría ser interesante comprobar si, pese a ello, una red es capaz de extraer valor de los datos añadidos.

Durante el entrenamiento de los modelos, se observaba una fuerte irregularidad en la evolución del *loss* y el *accuracy* de validación. En las ilustraciones 10, 11, 14 y 17 se observa cómo no se da una curva suave, sino que evoluciona a base de saltos bruscos e impredecibles. Esto podría indicar que las distintas clases se encuentran muy cercanas entre sí, y al modelo le cuesta encontrar el detalle suficiente para separarlas de forma consistente. En la resolución de este problema, el uso de imágenes con mayor resolución podría ayudar notablemente.

De hecho, en la base de datos del ADNI ya hay disponibles imágenes tomadas con imanes de 3T (teslas), aunque son mucho menos numerosas que las imágenes de 1.5T utilizadas en este trabajo. Esta medida, los teslas, representan básicamente la resolución con la que se pueden obtener las imágenes, y los investigadores están impulsando escáneres de 7T, 10T y hasta 21.1T, aunque no todos ellos están listos para su despliegue en entornos reales (Nowogrodzki, 2018). Por tanto, este trabajo podría actualizarse de forma recurrente según la resolución de las imágenes va aumentando. Es importante tener en cuenta que una

mayor resolución proporciona mayor detalle, pero también hace el entrenamiento mucho más costoso.

Por último, también se deben tener en cuenta los avances en *Deep Learning* en general a la hora de abrir nuevas vías de investigación para esta línea de trabajo. Constantemente se publican nuevas técnicas o arquitecturas que mejoran a las anteriores, y que permiten el entrenamiento de redes más profundas, una convergencia más rápida, mejores resultados, etc. Por ejemplo, las técnica de *Batch Normalization* supuso una mejora muy notable en el entrenamiento de redes profundas (Ioffe y Szegedy, 2015). Próximamente podrían surgir técnicas que supongan un avance similar. Incluso ya se han dado ciertos avances que aún no se han llevado al campo del diagnóstico del Alzheimer, como las redes Inception-ResNet (Szegedy, Ioffe, Vanhoucke, y Alemi, 2016) o la nueva función de activación *e-swish* (Alcaide, 2018).

5 Referencias

2018 reform of EU data protection rules. (2018). Recuperado de https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Zheng, X. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Recuperado de https://www.tensorflow.org/

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann Machines. *Cognitive Science*, *9*(1), 147-169.

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L., & Erickson, B. J. (2017). Deep learning for brain MRI segmentation: state of the art and future directions. *Journal of digital imaging*, *30*(4), 449-459.

Alcaide, E. (2018). E-swish: Adjusting Activations to Different Network Depths, 1-13. Recuperado de http://arxiv.org/abs/1801.07145

Beare, R., Lowekamp, B., & Yaniv, Z. (2018). Image Segmentation, Registration and Characterization in R with SimpleITK. *Journal of statistical software*, *86*.

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. En *Advances in neural information processing systems* (pp. 153-160).

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(Feb), 281-305.

Castro, E., Ulloa, A., Plis, S. M., Turner, J. A., & Calhoun, V. D. (2015). Generation of synthetic structural magnetic resonance images for deep learning pre-training. En *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)* (pp. 1057-1060). https://doi.org/10.1109/ISBI.2015.7164053

Chen, H., Dou, Q., Yu, L., & Heng, P.-A. (2016). VoxResNet: Deep Voxelwise Residual Networks for Volumetric Brain Segmentation. *arXiv preprint arXiv:1608.05895*. Recuperado de http://arxiv.org/abs/1608.05895

Chen, M., Shi, X., Zhang, Y., Wu, D., & Guizani, M. (2017). Deep features learning for medical image analysis with convolutional autoencoder neural network. *IEEE Transactions on Big Data*.

Chollet, F., & others. (2015). Keras.

Chopra, P., & Yadav, S. K. (2017). Restricted Boltzmann machine and softmax regression for fault detection and classification. *Complex & Intelligent Systems*, *4*(1), 67-77. https://doi.org/10.1007/s40747-017-0054-8

Ciompi, F., de Hoop, B., van Riel, S. J., Chung, K., Scholten, E. T., Oudkerk, M., … van Ginneken, B. (2015). Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. *Medical image analysis*, *26*(1), 195-202.

Colaboratory: Frequently asked questions. (2018).

Collins, D. L., Neelin, P., Peters, T. M., & Evans, A. C. (1994). Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *Journal of computer assisted tomography*, *18*(2), 192-205.

Dilsizian, S. E., & Siegel, E. L. (2014). Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Current cardiology reports*, *16*(1), 441.

Ding, Y., Sohn, J. H., Kawczynski, M. G., Trivedi, H., Harnish, R., Jenkins, N. W., … others. (2018). A Deep learning model to predict a diagnosis of alzheimer disease by using 18F-FDG PET of the brain. *Radiology*, *290*(2), 456-464.

dltk.networks.regression_classification package. (2017). Recuperado de https://dltk.github.io/DLTK/api/dltk.networks.regression_classification.html#module-dltk.networks.regression_classification.resnet

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115. Recuperado de https://www.nature.com/articles/nature21056.epdf?author_access_token=8oxIcYWf5UNrNpHsUHd2StRgN0jAjWel9jnR3ZoTv0NXpMHRAJy8Qn10ys2O4tuPakXos4UhQAFZ750CsBNMMsISFHIKinKDMKjShCpHIlYPYUHhNzkn6pSnOCt0Ftf6

Evans, A. C. (1992). An MRI-based stereotactic atlas from 250 young normal subjects. *Soc. neurosci. abstr, 1992*.

Evans, A. C., Collins, D. L., Mills, S. R., Brown, E. D., Kelly, R. L., & Peters, T. M. (1993). 3D statistical neuroanatomical models from 305 MRI volumes. En *1993 IEEE conference*

record nuclear science symposium and medical imaging conference (pp. 1813-1817).

Evans, A. C., Marrett, S., Neelin, P., Collins, L., Worsley, K., Dai, W., … Bub, D. (1992). Anatomical mapping of functional activation in stereotactic coordinate space. *Neuroimage*, *1*(1), 43-53.

Fang, C., Li, C., Cabrerizo, M., Barreto, A., Andrian, J., Loewenstein, D., … Adjouadi, M. (2017). A Novel Gaussian Discriminant Analysis-based Computer Aided Diagnosis System for Screening Different Stages of Alzheimer's Disease. En *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)* (pp. 279-284).

Feldman, M. D. (2004). Positron Emission Tomography (PET) for the Evaluation of Alzheimer's Disease/Dementia. *California Technology Assessment Forum*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative Adversarial Nets. *Advances in neural information processing systems*, 2672-2680. https://doi.org/10.1016/B978-0-408-00109-0.50001-8

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in neuroinformatics*, *5*, 13.

Greenspan, H., van Ginneken, B., & Summers, R. M. (2016). Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. *IEEE Transactions on Medical Imaging*, *35*(5), 1153-1159. https://doi.org/10.1109/tmi.2016.2553401

Gretzel, U., Sigala, M., Xiang, Z., & Koo, C. (2015). Smart tourism: foundations and developments. *Electronic Markets*, *25*(3), 179-188.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., … Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA - Journal of the American Medical Association*, *316*(22), 2402-2410. https://doi.org/10.1001/jama.2016.17216

Guo, X., Liu, X., Zhu, E., & Yin, J. (2017). Deep Clustering with Convolutional Autoencoder. En *International Conference on Neural Information Processing* (pp. 373-382). https://doi.org/10.1299/jsmemag.90.823_758

Gupta, A., Ayhan, M., & Maida, A. (2013). Natural image bases to represent neuroimaging data. En *International conference on machine learning* (pp. 987-994).

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. Recuperado de http://arxiv.org/abs/1512.03385

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. En *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Hellström, E. (2017). Feature learning with deep neural networks for keystroke biometrics. *Thesis*, 75.

Hellström, E. (2018). Feature learning with deep neural networks for keystroke biometrics: A study of supervised pre-training and autoencoders.

Hinton, G., Osindero, S., & Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, *18*(7), 1527-1554. https://doi.org/10.1162/neco.2006.18.7.1527

Hoffman, J. M., Welsh-Bohmer, K. A., Hanson, M., Crain, B., Hulette, C., Earl, N., & Coleman, R. E. (2000). FDG PET imaging in patients with pathologically verified dementia. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, *41*(11), 1920-1928. Recuperado de http://www.ncbi.nlm.nih.gov/pubmed/11079505

Hosseini-Asl, E., Gimel'farb, G., & El-Baz, A. (2016). Alzheimer's disease diagnostics by a deeply supervised adaptable 3D convolutional network. *arXiv preprint arXiv:1607.00556*.

Hosseini-Asl, E., Keynton, R., & El-Baz, A. (2016). Alzheimer's disease diagnostics by adaptation of 3D convolutional network. En *2016 IEEE International Conference on Image Processing (ICIP)* (pp. 126-130).

Imabayashi, E., Matsuda, H., Asada, T., Ohnishi, T., Sakamoto, S., Nakano, S., & Inoue, T. (2004). Superiority of 3-dimensional stereotactic surface projection analysis over visual inspection in discrimination of patients with very early Alzheimer's disease from controls using brain perfusion SPECT. *Journal of Nuclear Medicine*, *45*(9), 1450-1457.

Imagenet. (2016). Recuperado de http://www.image-net.org/

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Recuperado de http://arxiv.org/abs/1502.03167

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., … Weiner, M. W. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI

methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *27*(4), 685-691. https://doi.org/10.1002/jmri.21049

JihongJu. (2017). keras-resnet3d. Recuperado de https://github.com/JihongJu/keras-resnet3d

Jones, E., Oliphant, T., Peterson, P., & others. (2001). SciPy: Open source scientific tools for Python. Recuperado de http://www.scipy.org/

Ker, J., Wang, L., Rao, J., & Lim, T. (2018). Deep Learning Applications in Medical Image Analysis. *IEEE Access*, *6*, 9375-9389. https://doi.org/10.1109/ACCESS.2017.2788044

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klein, A., Andersson, J., Ardekani, B. A., Ashburner, J., Avants, B., Chiang, M.-C., … others. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage*, *46*(3), 786-802.

Klöppel, S., Stonnington, C. M., Barnes, J., Chen, F., Chu, C., Good, C. D., … others. (2008). Accuracy of dementia diagnosis—a direct comparison between radiologists and a computerized method. *Brain*, *131*(11), 2969-2974.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., … Willing, C. (2016). Jupyter Notebooks -- a publishing format for reproducible computational workflows. En F. Loizides & B. Schmidt (Eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (pp. 87-90).

Korolev, S., Safiullin, A., Belyaev, M., & Dodonova, Y. (2017). Residual and plain convolutional neural networks for 3D brain MRI classification. *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, 835-838.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. En *Advances in neural information processing systems* (pp. 1097-1105).

Lakhani, P., Gray, D. L., Pett, C. R., Nagy, P., & Shih, G. (2018). Hello world deep learning in medical imaging. *Journal of digital imaging*, *31*(3), 283-289.

Large Scale Visual Recognition Challenge. (2015). Recuperado de http://www.image-net.org/challenges/LSVRC/

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., & others. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.

Li, F., Tran, L., Thung, K.-H., Ji, S., Shen, D., & Li, J. (2015). A robust deep model for improved classification of AD/MCI patients. *IEEE journal of biomedical and health informatics*, *19*(5), 1610-1616.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., … Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, *42*(1995), 60-88. https://doi.org/10.1016/j.media.2017.07.005

Liu, Manhua, Zhang, D., Shen, D., & Initiative, the A. D. N. (2012). Ensemble Sparse Classification of Alzheimer's Disease. *NeuroImage*, *60*(2), 1106-1116. https://doi.org/10.1016/j.pestbp.2011.02.012.Investigations

Liu, Mingxia, Zhang, D., Adeli, E., & Shen, D. (2016). Inherent structure-based multiview learning with multitemplate feature representation for alzheimer's disease diagnosis. *IEEE Transactions on Biomedical Engineering*, *63*(7), 1473-1482.

Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., … others. (2015). Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. *IEEE Transactions on Biomedical Engineering*, *62*(4), 1132-1140.

Lo, S.-C., Lou, S.-L., Lin, J.-S., Freedman, M. T., Chien, M. V, & Mun, S. K. (1995). Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Transactions on Medical Imaging*, *14*(4), 711-718.

Lowekamp, B. C., Chen, D. T., Ibáñez, L., & Blezek, D. (2013). The design of SimpleITK. *Frontiers in neuroinformatics*, *7*, 45.

Marstal, K., Berendsen, F., Staring, M., & Klein, S. (2016). SimpleElastix: A user-friendly, multi-lingual library for medical image registration. En *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 134-142).

Masci, J., Meier, U., Cire\csan, D., & Schmidhuber, J. (2011). Stacked convolutional auto-encoders for hierarchical feature extraction. En *International Conference on Artificial Neural Networks* (pp. 52-59).

Matsuda, H. (2007). Role of neuroimaging in Alzheimer's disease, with emphasis on brain perfusion SPECT. *Journal of Nuclear Medicine*, *48*(8), 1289-1300.

Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., & Tourassi, G. D.

(2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural networks*, *21*(2-3), 427-436.

Millington, I., & Funge, J. (2009). *Artificial intelligence for games*. CRC Press.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, *65*, 211-222.

Nordberg, A., Rinne, J. O., Kadir, A., & Lngström, B. (2010). The use of PET in Alzheimer disease. *Nature Reviews Neurology*, *6*(2), 78-87. https://doi.org/10.1038/nrneurol.2009.217

Nowogrodzki, A. (2018). The world's strongest MRI machines are pushing human imaging to new limits.

O'Connor, J. P. B., Aboagye, E. O., Adams, J. E., Aerts, H. J. W. L., Barrington, S. F., Beer, A. J., … Waterton, J. C. (2017). Imaging biomarker roadmap for cancer studies. *Nature reviews. Clinical oncology*, *14*(3), 169-186. https://doi.org/10.1038/nrclinonc.2016.162

OASIS Brains. (2018). *Oasis-3 : Imaging Methods and Data Dictionary*.

OASIS Brains Datasets. (s. f.). Recuperado de https://www.oasis-brains.org/#data

Pawlowski, N., S. Ira Ktena, Lee, M. C. H., Kainz, B., Rueckert, D., Glocker, B., & Rajchl, M. (2017). DLTK: State of the Art Reference Implementations for Deep Learning on Medical Images. *arXiv preprint arXiv:1711.06853*.

Payan, A., & Montana, G. (2015). Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. *arXiv preprint arXiv:1502.02506*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, *12*, 2825-2830.

Pichler, B. J., Kolb, A., Nagele, T., & Schlemmer, H.-P. (2010). PET/MRI: Paving the Way for the Next Generation of Clinical Multimodality Imaging Applications. *Journal of Nuclear Medicine*, *51*(3), 333-336. https://doi.org/10.2967/jnumed.109.061853

Rajchl, M., Ktena, S. I., & Pawlowski, N. (2018). An Introduction to Biomedical Image Analysis with TensorFlow and DLTK. Recuperado 9 de junio de 2019, de https://medium.com/tensorflow/an-introduction-to-biomedical-image-analysis-with-

tensorflow-and-dltk-2c25304e7c13

Razzak, M. I., Naz, S., & Zaib, A. (2018). Deep Learning for Medical Image Processing: Overview, Challenges and the Future. En *Classification in BioApps* (pp. 323-350). https://doi.org/10.1007/978-3-319-65981-7_12

Real, E., Aggarwal, A., Huang, Y., & Le, Q. V. (2018). Regularized Evolution for Image Classifier Architecture Search. *arXiv preprint arXiv:1802.01548*, 52-54. Recuperado de http://arxiv.org/abs/1802.01548

Reitz, C., Brayne, C., & Mayeux, R. (2011). Epidemiology of Alzheimer disease. *Nature Reviews Neurology*, *7*(3), 137.

Rueckert, D., Sonoda, L. I., Hayes, C., Hill, D. L. G., Leach, M. O., & Hawkes, D. J. (1999). Nonrigid registration using free-form deformations: application to breast MR images. *IEEE transactions on medical imaging*, *18*(8), 712-721.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

Salakhutdinov, R., & Hinton, G. (2009). Deep Boltzmann Machines. *Artificial intelligence and statistics*, 448-455. Recuperado de http://jmlr.org/proceedings/papers/v5/salakhutdinov09a/salakhutdinov09a.pdf

Salakhutdinov, R., Mnih, A., & Hinton, G. (2007). Restricted Boltzmann Machines for Collaborative Filtering. *Proceedings of the International Conference on Ma-chine Learning*, *24*, 791–798.

Sarraf, S., Tofighi, G., & others. (2016). DeepAD: Alzheimer' s disease classification via deep convolutional neural networks using MRI and fMRI. *BioRxiv*, 70441.

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual review of biomedical engineering*, *19*, 221-248.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smith, S. M. (2002). Fast robust automated brain extraction. *Human brain mapping*, *17*(3), 143-155.

Strimbu, K., & Tavel, J. A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS*, *5*(6), 463. https://doi.org/10.1097/COH.0b013e32833ed177.What

Suk, H.-I., Lee, S.-W., Shen, D., Initiative, A. D. N., & others. (2014). Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage*, *101*, 569-582.

Suk, H.-I., Lee, S.-W., Shen, D., Initiative, A. D. N., & others. (2015). Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Structure and Function*, *220*(2), 841-859.

Suk, H. Il, & Shen, D. (2013). Deep learning-based feature representation for AD/MCI classification. En *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 583-590). https://doi.org/10.1007/978-3-642-40763-5_72

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2016). Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. Recuperado de http://arxiv.org/abs/1602.07261

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. En *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. En *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).

tf.image.per_image_standardization. (s. f.). Recuperado de https://www.tensorflow.org/api_docs/python/tf/image/per_image_standardization

Thies, W., & Bleiler, L. (2012). 2012 Alzheimer's disease facts and figures. *Alzheimer's & dementia: the journal of the Alzheimer's Association*.

Tong, T., Wolz, R., Gao, Q., Guerrero, R., Hajnal, J. V, Rueckert, D., … others. (2014). Multiple instance learning for classification of dementia in brain MRI. *Medical image analysis*, *18*(5), 808-818.

Using TFRecords and tf.Example. (s. f.). Recuperado de https://www.tensorflow.org/tutorials/load_data/tf_records

Valetts, L. M., Navarro, S. B., & Gesa, R. F. (2008). Modelling Collaborative Competence Level Using Machine Learning Techniques. En *e-Learning* (pp. 56-60).

Vesal, S., Ravikumar, N., Davari, A. A., Ellmann, S., & Maier, A. (2017). Classification of Breast Cancer Histology Images Using Transfer Learning. *PloS one*, *12*(6), 812-819.

https://doi.org/10.1007/978-3-319-93000-8_92

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P.-A. (2010). Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*, *11*(3), 3371-3408. https://doi.org/10.1111/1467-8535.00290

Woods, R. P., Mazziotta, J. C., Cherry, S. R., & others. (1993). MRI-PET registration with automated algorithm. *Journal of computer assisted tomography*, *17*, 536.

Wu, R., Yan, S., Shan, Y., Dang, Q., & Sun, G. (2015). Deep Image: Scaling up Image Recognition. Recuperado de http://arxiv.org/abs/1501.02876

Yang, W., Lui, R. L. M., Gao, J.-H., Chan, T. F., Yau, S.-T., Sperling, R. A., & Huang, X. (2011). Independent component analysis-based classification of Alzheimer's disease MRI data. *Journal of Alzheimer's disease*, *24*(4), 775-783.

Yaniv, Z., Lowekamp, B. C., Johnson, H. J., & Beare, R. (2018). SimpleITK image-analysis notebooks: a collaborative environment for education and reproducible research. *Journal of digital imaging*, *31*(3), 290-303.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. En *European conference on computer vision* (pp. 818-833).

Zu, C., Jie, B., Liu, M., Chen, S., Shen, D., Zhang, D., … others. (2016). Label-aligned multi-task feature learning for multimodal classification of Alzheimer's disease and mild cognitive impairment. *Brain imaging and behavior*, *10*(4), 1148-1159.