Universidad de Valladolid

FACULTAD DE MEDICINA
ESCUELA DE INGENIERÍAS INDUSTRIALES

TRABAJO FIN DE GRADO

GRADO EN INGENIERÍA BIOMÉDICA

MRI sintética mediante Deep Learning para mejorar el diagnóstico de tumores cerebrales

Autor:

Jorge Rueda Ramos

Tutores:

Elisa Moya Sáez y Carlos Alberola López

Valladolid, 30 de septiembre de 2024

Título:	**MRI sintética mediante Deep Learning para mejorar el diagnóstico de tumores cerebrales**
Autor:	**Jorge Rueda Ramos**
Tutores:	**Elisa Moya Sáez y Carlos Alberola López**
Departamento:	**TSCIT**

TRIBUNAL

Presidente:	**Santiago Aja Fernández**
Secretario:	**Rodrigo de Luis García**
Vocal:	**Marcos Antonio Martín Fernández**
Suplente:	**Rosa María Menchón Lara**
Suplente:	**Miguel Ángel Martín Fernández**

Fecha:

Calificación:

RESUMEN DEL PROYECTO

En este Trabajo de Fin de Grado se aborda la exploración y aplicación de técnicas de aprendizaje profundo que permitan sintetizar imágenes ponderadas post-contraste, sin la utilización de agente de contraste, a partir de mapas paramétricos cuantitativos de pacientes con distintos grados de gliomas. Este campo de investigación posee importantes motivaciones como son los peligros asociados a los agentes de contraste basados en gadolinio, o las limitaciones que presenta el diagnóstico tradicional por la escala arbitraria de las imágenes ponderadas, entre otras.

Para ello, se tomaron como punto de partida los mapas paramétricos, propiedades magnéticas tisulares que incluyen el tiempo de relajación longitudinal (T1), el tiempo de relajación transversal (T2), y la densidad de protones (DP). Gracias a la naturaleza cuantitativa inherente (imagen por resonancia magnética cuantitativa) de ellos, se dispone de un conjunto de datos más robusto para aplicar las técnicas desarrolladas en este trabajo.

En el ámbito del aprendizaje profundo, el trabajo gira en torno a dos arquitecturas de red neuronal, dos enfoques de aprendizaje automático, y dos configuraciones de datos de entrada a nuestros modelos. Las redes utilizadas son una red neuronal convolucional con arquitectura U-Net y un transformador de visión aplicado a la arquitectura original (U-Net ViT), para cada cual se aplican dos tipos de aprendizaje, supervisado y autosupervisado, con dos estructuras de datos de entrada cada uno, la imagen completa (cortes completos), y basada en parches. Además, se ha aplicado la técnica de validación cruzada dejando uno fuera (*leave-one-out, LOO*, con la que comparar cada uno de estos experimentos.

Los modelos nos devuelven predicciones de la imagen ponderada en T1 post-contraste a partir de mapas obtenidos sin la aplicación de gadolinio. Las imágenes obtenidas con aprendizaje profundo se comparan con las imágenes adquiridas, y sobre ellas se calculan métricas de calidad a raíz de las cuales se realizan las comparaciones y discusiones finales entre experimentos. Estas métricas son el error cuadrático medio (*mean squared error*, MSE), el índice de similitud estructural (*structural similarity index measure*, SSIM), y la proporción máxima de señal a ruido (*Peak Signal-to-Noise Ratio*, PSNR).

PALABRAS CLAVE

IRM sintética, aprendizaje profundo, realce, agentes de contraste, tumores cerebrales

ABSTRACT

In this thesis we explore and apply new deep learning techniques to synthesise post-contrast weighted images, without the use of contrast agents, from quantitative parametric maps of patients with different grades of gliomas. This field of research has important motivations such as the dangers associated with gadolinium-based contrast agents, or the limitations of traditional diagnosis due to the arbitrary scale of the weighted images, among many others.

For this purpose, parametric maps, tissue magnetic properties including longitudinal relaxation time (T1), transverse relaxation time (T2), and proton density (PD), were taken as a starting point. Due to their inherent quantitative nature (quantitative magnetic resonance imaging), a more robust and manageable dataset is available to apply the techniques developed in this work.

In the deep learning domain, the thesis revolves around two neural network architectures, two types of machine learning, and two configurations of input data to our models. The networks used are a convolutional neural network with U-Net architecture and a vision transformer applied to the original arquitecture (U-Net ViT) , for each of which two types of learning are applied, supervised and self-supervised, with two input data structures each, full image (full slices), and patch-based. In addition, the leave-one-out (LOO) cross-validation technique has been applied to compare each of these experiments.

The models return post-contrast T1-weighted image predictions from maps obtained without gadolinium application. The images obtained with deep learning are compared with the acquired images, and quality metrics are calculated on the basis of which the final comparisons and discussions between experiments are made. These metrics are the mean squared error (MSE), the structural similarity index measure (SSIM), and the peak signal-to-noise ratio (PSNR).

KEYWORDS

Synthetic MRI, deep learning, enhancement, contrast agents, brain tumors

Agradecimientos

Quisiera empezar estas líneas expresando mi profunda gratitud hacia mis tutores, Elisa Moya Sáez y Carlos Alberola López, por su apoyo, orientación, y constante disposición en todo lo relacionado con mi trabajo. A lo largo de esta travesía he enfrentado momentos buenos y malos, muchos retos de gran dificultad para mí, que se han visto acrecentados por mi forma de ser. Gracias, Elisa, porque tu actitud serena y tenaz para manejar estas adversidades me ha servido de inspiración, y puedo asegurar que esta es la lección más importante que me llevo de todo este tiempo. Gracias por darme la oportunidad de explorar un tema tan pionero e interesante, y haberme acogido tan cálidamente desde los inicios. Gracias por la libertad y confianza otorgadas.

De igual modo, agradezco la simpatía y honestidad de todos los profesionales que trabajan día a día en el entorno de trabajo donde me he desenvuelto, el Laboratorio de Procesado de Imagen. Muy especialmente a aquellos que han decidido conocerme y darme una oportunidad. La atmósfera de unidad y colaboración que se respira en el "Lab" se debe a cada uno de vosotros.

Por último, agradecer a familiares y amigos la paciencia y confianza depositada en mí, especialmente a mi abuela Antonia. Sin ella, nunca habría tomado la decisión de embarcarme en este proyecto, por el tema en el que se fundamenta.

Índice general

Índice general vii

1. Introducción 1

1.1. Motivación . . . 1
1.2. Objetivos . . . 4
1.3. Metodología . . . 5
1.4. Medios materiales . . . 6
1.5. Estructura del documento . . . 8

2. Estado del arte 9

2.1. Imagen por Resonancia Magnética . . . 9
2.1.1. Contexto . . . 9
2.1.2. Principios físicos . . . 9
2.1.3. Formación de la imagen . . . 13
2.1.4. Secuencias típicas . . . 14
2.1.5. Ecuaciones teóricas de las secuencias de pulsos . . . 16
2.2. Agentes de contraste . . . 17
2.3. Aprendizaje profundo . . . 19
2.3.1. Redes neuronales artificiales . . . 19
2.3.2. Tipos de aprendizaje . . . 21
2.3.3. Redes neuronales convolucionales . . . 22
2.3.4. Arquitectura U-Net . . . 24
2.3.5. Transformadores de visión . . . 25

3. Métodos 27

3.1. Preprocesamiento de la base de datos . . . 27
3.2. Construcción de entradas y salidas del modelo . . . 28
3.2.1. Imagen completa . . . 30

3.2.2. Basado en parches . . . 30
3.2.3. Aumento de datos de parches con realce . . . 32
3.3. Arquitecturas de redes neuronales aplicadas . . . 32
3.3.1. Arquitectura U-Net . . . 33
3.3.2. Arquitectura U-Net ViT . . . 36
3.4. Tipos de aprendizaje . . . 38
3.4.1. Aprendizaje supervisado . . . 38
3.4.2. Aprendizaje autosupervisado . . . 39
3.5. Reconstrucción de las parches . . . 40
3.6. Experimentos . . . 40
3.7. Evaluación . . . 42

4. Resultados ***45***
4.1. Por exposición visual . . . 45
4.2. Por métricas de calidad . . . 51

5. Conclusiones ***57***
5.1. Conclusiones . . . 57
5.2. Líneas futuras . . . 59

Bibliografía ***61***

A. Parámetros de adquisición de la base de datos ***67***

Capítulo 1

Introducción

1.1 Motivación

El glioma maligno se postula, dentro de los tumores cerebrales primarios en adultos, como el tipo más común de ellos (60 % de todos los tumores cerebrales primarios) [1]. Posee una tasa de incidencia media anual de en torno a 5.26 casos por cada 100000 habitantes, siendo más frecuente su aparición entre la sexta y la octava década de vida [1]. Debido a ello, y con el envejecimiento de la población, se espera un incremento en el número de pacientes con esta enfermedad durante los próximos años.

El origen de estos tumores malignos tiene lugar en las células de la glia del sistema nervioso central (SNC), células de soporte y protección del organismo. La variedad de gliomas se corresponde con los distintos tipos de células de la glia existentes, pues se originan a partir de un tipo específico de estas. También resulta relevante destacar que, aunque el término "malignidad" en otras partes del organismo adquiere una connotación de propensión a invadir y metastatizar, en lo referido al SNC, este término se asocia a la agresividad global del tumor, pues no viajan ni se instalan más allá del sistema nervioso (por lo general). A pesar de ello, la familia de gliomas comparte, en su conjunto, uno de los pronósticos más desoladores de todas las neoplasias del organismo [2]. En especial, el glioblastoma, con una supervivencia tras el diagnóstico del 40 % en el primer año, y de un 17 % en el segundo, aproximadamente [3].

En resumen, el reto biomédico y social que enfrentamos presenta una especial relevancia por el aumento de casos naturales a causa del envejecimiento de la sociedad occidental actual, y por la preocupación y gravedad que suscita el pronóstico tan desfavorable que presentan este

tipo de tumores. Por todo ello, será vital la investigación en la detección de estas neoplasias, que permitan programar correctamente el tratamiento y la cirugía a realizar. Concretamente, el trabajo se centra en la síntesis de imágenes post-contraste a partir de las cuales podremos identificar el realce tumoral, clave en la planificación de la resección tumoral, y muy informativo de la agresividad de tumores de alto grado [4].

Atendiendo al diagnóstico de esta enfermedad, la principal modalidad de imagen médica utilizada es la resonancia magnética (RM). Esto es debido a su carácter no invasivo e inocuo, que la distingue de otras modalidades como la tomografía computarizada (TC), que aplica radiación ionizante. Además, presenta muy buenos contrastes entre los tejidos blandos, lo que la dota de una especial relevancia en este campo. Sin embargo, también posee importantes limitaciones como son los largos tiempos de adquisición en base a los protocolos utilizados, el coste de los equipos y de las pruebas, o las distintas contraindicaciones de origen ferromagnético y atribuibles al paciente (marcapasos, prótesis, ...), que pueden restringir su uso.

La imagen por resonancia magnética es una técnica de imagen basada en las propiedades magnéticas inherentes a las moléculas del organismo. Típicamente, las adquisiciones son de carácter cualitativo (RM cualitativa o ponderada), y se obtienen de la aplicación de unas secuencias de pulsos, y de una elección de parámetros de adquisición. La selección de las secuencias y los parámetros origina una ponderación específica de las propiedades magnéticas tisulares. Cada una de las ponderaciones de imagen adquiridas (y utilizadas en el proyecto), va a proporcionar información adicional y complementaria al diagnóstico, ya que las estructuras y patologías presentes serán más o menos visibles, en función de la ponderación en cuestión. El problema de las imágenes ponderadas es que poseen una escala arbitraria que dificulta el uso de métodos cuantitativos de diagnóstico.

Otra modalidad a tener en cuenta es la RM cuantitativa. Esta, originalmente, precisaba de al menos dos imágenes cualitativas (idealmente más) para poder cuantificar correctamente las propiedades tisulares, concluyendo en un tiempo de adquisición mucho mayor en comparación con la RM ponderada, y, en definitiva, en una modalidad normalmente impracticable en la clínica real [5]. Este importante cuello de botella llevó a la introducción de avances novedosos en el ámbito de la relaxometría (la técnica tradicional de RM cuantitativa) y, posteriormente, a la

inclusión de un nuevo procedimiento de adquisición rápido y eficiente en el tiempo: la compilación de imágenes por resonancia magnética (MAGiC) [6]. En una última etapa de este método, se realizó un refinamiento adicional que logró combinar en una sola adquisición de imagen de RM los tiempos de relajación longitudinal y transversal (T1 y T2, respectivamente), el mapa de densidad de protones, DP, y la generación de las imágenes cualitativas ponderada en T1 (T1w), ponderada en T2 (T2w) y recuperación de inversión atenuada por fluido ponderada en T2 (T2w-FLAIR, de sus siglas en inglés *Fluid Attenuated Inversion Recovery*), a partir de los mapas [6].

Nuestro conjunto de datos estará conformado por la imagen ponderada T1w, la ponderada T2w, la T2w-FLAIR, y la ponderación en T1 posterior a la inyección de un agente de contraste (post-T1w), todas ellas obtenidas a partir del protocolo estándar para la evaluación de gliomas [7,8]. A mayores, tendremos como registros cuantitativos los mapas paramétricos obtenidos con MAGiC (*magnetic resonance imaging compilation*): T1, T2, y DP.

Es importante recabar en la imagen post-T1w, ya que la propia naturaleza de los agentes de contraste implica una serie de inconvenientes y desafíos. La familia de agentes de contraste típicamente utilizada en RM son los basados en gadolinio (*Gadolinium-Based Contrast Agents*, GBCA), y los problemas asociados se definen a continuación:

- Reacciones adversas de los pacientes [9].
- Situaciones aparatosas e incómodas para el paciente, durante la inyección intravenosa.
- Riesgo de retención (almacenamiento) de GBCA en los tejidos del cerebro [10].
- Impacto medioambiental: contaminación de aguas superficiales, aguas potables, sedimentos, e incluso de organismos vivos [11].

Este trabajo tendrá como unidad funcional el aprendizaje profundo, un enfoque dentro del aprendizaje automático basado en las redes neuronales artificiales, e inspirado en el funcionamiento del cerebro humano. Las redes neuronales son especialmente interesantes debido a su estructura flexible, que permite su modificación y adaptación a una amplia variedad de contextos en torno a las distintas técnicas de aprendizaje en las que se pueden emplear [12].

El método empleado en el trabajo se basa en la aplicación de técnicas de aprendizaje profundo para la síntesis de imagen. Aunque estas técnicas ofrecen la ventaja de no aumentar el tiempo de las pruebas y son cómodas de utilizar, presentan una limitación significativa: la falta de conjuntos de datos con los que poder explotarlas.

Gracias a los registros cuantitativos con los que contamos, buscaremos predecir mediante modelos de aprendizaje profundo la imagen post-T1w, sin la utilización de gadolinio. En concreto, haremos especial énfasis en la detección del realce propio de estas imágenes post-contraste. Con ello, tratamos de resolver el problema biomédico, como se estableció en la primera parte de la motivación del trabajo, y a su vez tratamos de implementar una forma puntera de evitar los riesgos del GBCA por su uso en la propia rutina de diagnóstico y tratamiento de estos pacientes. Del mismo modo, con la aplicación del aprendizaje profundo buscaremos obtener imágenes sintéticas, sin aumentar los tiempos de adquisición.

1.2 Objetivos

El objetivo principal de este Trabajo de Fin de Grado es **mejorar el diagnóstico de los tumores cerebrales mediante la exploración de diversas configuraciones de aprendizaje profundo, evaluando arquitecturas, métodos de entrenamiento y entradas, con el fin de sintetizar imágenes post-T1w sin el uso de agentes de contraste, a partir de los mapas paramétricos T1, T2 y DP**.

A su vez, podemos desarrollar el objetivo principal en una serie de subobjetivos:

- Evitar el uso de GBCA y sus efectos adversos, gracias a la obtención sintética de las imágenes post-contraste.
- Detectar el realce tumoral de la imagen post-T1w, propio de la ruptura de la barrera hematoencefálica.
- Ratificar que, por medio de técnicas de aprendizaje profundo, el método de obtención de estas imágenes sintéticas no añade tiempo adicional al ya dilatado tiempo de adquisición de los protocolos de las secuencias rutinarias.

1.3 Metodología

La metodología que se ha empleado en este Trabajo de Fin de Grado se descompone en las siguientes etapas, cada una de las cuales se subdivide en varias fases (tareas asociadas) que completan la consecución total de la etapa en cuestión.

1. **Etapa de capacitación**:

   a) Adquisición de competencias básicas para el manejo del sistema operativo *Linux*.

   b) Adquisición de competencias básicas para el manejo del entorno de desarrollo integrado (IDE) *Pycharm* [13]. Creación del entorno virtual, configuración del intérprete, instalación de los paquetes pertinentes, y ajuste del entorno con los servidores del laboratorio de trabajo (tarjetas gráficas, ...).

   c) Desarrollo de competencias básicas para el manejo del *framework MRview* [14], software enfocado en el procesamiento y visualización de imágenes médicas.

2. **Etapa de formación**:

   a) Estudio de los fundamentos de la resonancia magnética, con un enfoque en las modalidades de imagen manipuladas en el trabajo.

   b) Formación detallada en aprendizaje profundo y las redes y avances recientes que presenta, en materia de la síntesis de imagen.

   c) Comprensión del código de Python otorgado por los tutores para la consecución del trabajo.

3. **Etapa de desarrollo**:

   a) Programación de los distintos experimentos siguiendo un mismo patrón: fichero de carga y preprocesamiento de los datos, fichero con el modelo de red neuronal, fichero de ejecución o arranque, fichero de guardado, y los ficheros con las funciones necesarias asociadas.

   b) Búsqueda detallada de estudios previos aplicables a mi trabajo, y adecuación de los mismos para la mejora de resultados.

4. **Etapa de evaluación**:

   a) Obtención de métricas de utilidad de cada experimento.

   b) Comparación de resultados entre los distintos experimentos y consecuente discusión.

   c) Reflexión final del trabajo y de las líneas futuras que nacen a partir de este.

1.4 Medios materiales

Los medios materiales utilizados en el desarrollo del trabajo en materia de *hardware* y *software* han sido:

Software

- Python 3.9 [15]: lenguaje de programación técnico que permite trabajar rápidamente e integrar sistemas de forma efectiva.
- Pycharm [13]: entorno de desarrollo integrado (IDE) de Python para la ciencia de datos y el desarrollo web.
- MRView de MRtrix3 [14]: herramienta de visualización de imágenes del paquete de software MRtrix3. Este paquete es una colección de herramientas utilizadas en el análisis y procesamiento de las imágenes por resonancia magnética.
- LaTeX [16]: sistema de composición de textos, orientado a la creación de documentos escritos de alta calidad tipográfica.
- Overleaf [17]: editor *online* de LaTeX donde se desarrolla la memoria del Trabajo de Fin de Grado.
- MATLAB [18]: plataforma de programación y cálculo numérico utilizada para la obtención de las figuras expuestas en el Capítulo 4.

Hardware. Se ha trabajado, con carácter general, en el PC de sobremesa del laboratorio y en los servidores del mismo. No obstante, también se han empleado muchas horas en el ordenador portátil personal, a tiempos en los servidores del laboratorio trabajando en remoto y, en otras ocasiones, con los propios recursos del ordenador. Se detalla a continuación:

- PC de sobremesa del LPI, Laboratorio de Procesado de Imagen, de la Universidad de Valladolid (UVa), con los siguientes componentes:

  - Procesador 16 × *AMD Ryzen* 7 5800X 8-Core Processor.
  - 32 GB de RAM.
  - Disco duro interno de 500 GB de capacidad.
  - Unidad de tarjeta gráfica Quadro RTX 4000.

- Ordenador portátil personal:
  - Procesador *Intel Core* i5-1035G1 CPU @ 1.00GHz 1.19 GHz
  - 16 GB de RAM.
  - Disco duro interno de 500 GB de capacidad.
  - Tarjeta gráfica *Intel(R)* UHD Graphics, integrada en la CPU.
- Servidores de trabajo disponibles en el LPI. Se ha trabajado mayoritariamente en el servidor isis, compuesto a su vez por 4 tarjetas gráficas de modelo *NVIDIA* RTX A5000, con una VRAM DE 24.564 GB cada una.

El conjunto de datos utilizado fue adquirido en el hospital *Erasmus Medical Center* de Róterdam (Países Bajos), en un equipo *3T Sigma Premier* de la compañía *General Electric Medical Systems*, entre los años 2018 y 2020. Las adquisiciones se realizaron con la aprobación de la Junta de Revisión Institucional y un consentimiento informado por escrito. Esta base de datos está compuesta por un total de 15 pacientes (6 mujeres y 9 hombres) diagnosticados con gliomas de distinto grado. La edad media de los 15 sujetos sigue una media de 39.33 años y una desviación típica de 10.40 años. Cabe destacar que previo a la adquisición de las imágenes, se procedió a la resección del tumor. A mayores, para uno de los 15 pacientes no se adquirió el registro T2w-FLAIR debido a una desviación del protocolo, y se ha prescindido del sujeto en su totalidad. Los detalles de adquisición de este banco de datos pueden ser consultados en la Tabla A.1, del Anexo A.

La base de datos está constituida por 4 modalidades de imagen (por paciente): T1w, T2w, T2w-FLAIR y la post-T1w, además de los mapas paramétricos pre-contraste T1, T2 y DP, obtenidos utilizando la técnica cuantitativa *MAGiC*. En el total de experimentos y pruebas realizadas, se ha trabajado con todos los tipos de registros mencionados en el anterior párrafo, a excepción del T2w-FLAIR, que no se ha empleado en ningún momento.

El espacio de trabajo ha sido el Laboratorio 26 Bis de la Escuela Técnica Superior de Ingenieros de Telecomunicación de la UVa.

1.5 Estructura del documento

La memoria del Trabajo de Fin de Grado se divide en 4 capítulos, introducidos a continuación:

En el capítulo 1 se detalla la motivación, los objetivos, la metodología, los medios materiales, y la propia estructura del documento (apartado actual). La motivación busca exponer el problema del que parte el trabajo, y las soluciones y mejoras que se han llevado a cabo. Además, se presenta de manera general información de los contenidos teóricos básicos en torno a los que gira este proyecto (imagen por resonancia magnética, agentes de contraste y aprendizaje profundo). En el apartado de los objetivos, se expone la meta a alcanzar en el trabajo y los subobjetivos específicos. La metodología la componen las distintas etapas en las que se ha desarrollado el trabajo, y los medios materiales son los recursos utilizados para la consecución del mismo.

En el capítulo 2 se establece el estado del arte del trabajo. Se profundiza en los contenidos expuestos en la introducción, y se desarrolla un contexto del que partir en los capítulos posteriores, especialmente, en el capítulo 3.

En el capítulo 3 se expone con detalle el método aplicado durante el trabajo para la consecución de los objetivos. En el transcurso del proyecto se han aplicado dos tipos de redes, una red neuronal convolucional con arquitectura *U-Net* y un transformador de visión en la arquitectura U-Net original (*U-Net ViT*), dos tipos de aprendizajes, supervisado y autosupervisado, y dos tipos de construcciones de entrada y salida al modelo, por imagen completa y basado en parches. Los distintos experimentos son resultado de combinarse estos aspectos.

En el capítulo 4 se presentan los experimentos y pruebas realizadas, así como las discusiones y comparaciones de los resultados a raíz del análisis de los mismos.

En el capítulo 5 se recogen las conclusiones generales extraídas con la elaboración de este Trabajo de Fin de Grado, así como las potenciales líneas de investigación y desarrollo a futuro, resultantes de su elaboración.

Capítulo 2

Estado del arte

2.1 Imagen por Resonancia Magnética

2.1.1 Contexto

La imagen por resonancia magnética (IRM) es una técnica de imagen médica basada en el fenómeno de la resonancia magnética nuclear (RMN), y caracterizada por aportar un gran detalle y unos altos contrastes en las imágenes [19]. Por tanto, se establece como una técnica de diagnóstico fundamental para el entorno clínico y sanitario.

Podemos distinguirla de muchas de las modalidades de imagen médica utilizadas en la actualidad, en que no utiliza radiación ionizante, y en consecuencia, su utilización no va a tener perjuicios en la salud del paciente tratado. Aunque podríamos pensar que por esta razón sería una técnica hegemónica en la sociedad actual, la modalidad de resonancia magnética también presenta muchas limitaciones, que restringen su uso.

2.1.2 Principios físicos

Profundizando en la física del fenómeno de la IRM, es decir, en los principios en los que se basa la formación de estas imágenes, debemos empezar por la RMN, y en particular por la unidad más fundamental de la materia, el átomo. El átomo es la unidad más pequeña en la que podemos dividir a la materia sin que esta pierda sus propiedades químicas, y está compuesto por un núcleo de protones (carga positiva) y neutrones (carga neutra), y una nube de electrones a su alrededor, de carga negativa. En el núcleo, cada protón va a girar sobre su propio eje, que

es lo que se conoce como momento angular del espín o simplemente espín [19]. Sin embargo, en la mayoría de casos, y sin presencia de un campo magnético externo, los protones y neutrones se alinean y su resultado deriva en una magnetización neta nula [20]. No obstante, esta magnetización neta nula no se da cuando el número de protones y neutrones es impar, siendo el principal exponente de esta condición en la RM el átomo de hidrógeno [20]. Este elemento adquiere especial relevancia en esta modalidad de imagen por su opulencia en el organismo, formando parte de las moléculas de agua de las diferentes estructuras humanas.

En el hidrógeno, el núcleo es poseedor de un único protón, el cual genera un pequeño campo magnético en el átomo (vector a lo largo del eje del espín), denominado momento magnético ($\vec{\mu}$), que estará relacionado con el momento angular del mismo espín ($\vec{J}$) a través de la relación o constante giromagnética (γ) propia de cada elemento [19]. La correspondencia de estos dos momentos viene dada por:

$$\vec{\mu} = \gamma \cdot \vec{J} \tag{2.1}$$

En lo referido a una muestra de hidrógenos cuando en ella se introduce un campo magnético estático externo, $\vec{B_0}$, la orientación de los espines de los distintos átomos será paralela o antiparalela al campo aplicado, es decir, a sus líneas de campo. Esto se debe al número cuántico de espín, que para el átomo de estudio es de $I = \frac{1}{2}$ (1 protón no emparejado, y 0 neutrones, por tanto, ningún neutrón no emparejado) [21]. El efecto de la aplicación de un campo magnético $\vec{B_0}$ en un sistema de núcleos de hidrógeno podemos visualizarlo en la Figura 2.1.

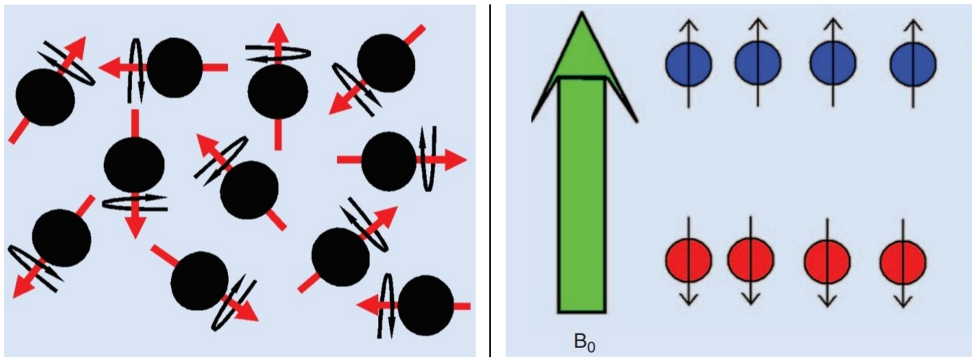


Figura 2.1: A la izquierda, orientación aleatoria de los espines en una situación normal. A la derecha, alineamiento de los espines en sus dos variantes, tras la aplicación de un campo magnético estático [19].

El estado preferido para cada átomo será el que le requiera de una menor energía para su colocación. Siendo así, idealmente, todos los espines se colocarían paralelos a $\vec{B_0}$, pero esto no sucede por los movimientos térmicos, que resultan en una transición de espines entre altos y bajos estados de energía (antiparalelo y paralelo al $\vec{B_0}$, respectivamente) [21]. No obstante, solo algunos se ubicarán en el estado de alta energía, y prevalecerá un predominio de espines en posición paralela, que harán que el vector de magnetización neta, $\vec{M}$, no sea nulo y que exista una magnetización neta positiva (suma de los momentos de cada protón de una muestra). La diferencia de energía de los dos estados será proporcional a la fuerza de $\vec{B_0}$ [21].

Tras la aplicación de $\vec{B_0}$, el momento magnético nuclear ($\vec{\mu}$) experimenta un torque, ($\vec{\tau}$), que tiende a alinearlo con el campo externo, siguiendo la relación [21]:

$$\vec{\tau} = \vec{\mu} \cdot \vec{B_0} \tag{2.2}$$

El par ejercido provoca un cambio en el momento angular del núcleo, causando precesión del momento magnético sobre la dirección de B_0 con carácter giroscópico, en vez de alinearse en su dirección [21]. Este fenómeno se conoce como precesión de Larmor, y la frecuencia angular a la que el núcleo lo hace sobre el campo aplicado, frecuencia de Larmor (ω_0). El signo menos de la ecuación que define la frecuencia ilustra la dirección de rotación, aunque es comúnmente ignorado.

$$\vec{\omega_0} = -\gamma \vec{B_0} \tag{2.3}$$

Hasta este momento, nos encontramos en un sistema al que se le ha aplicado un campo magnético estático $\vec{B_0}$, que ha orientado los espines en dirección paralela (más frecuente) y en dirección antiparalela, es decir, en el eje longitudinal. Esto es, que la práctica totalidad de $\vec{M}$ se producirá en el eje z (dirección de las líneas de campo de $\vec{B_0}$), mientras que en el plano xy el vector de magnetización será nulo. El problema radica en que no se puede medir una señal (por ende, obtener una imagen), si esta es paralela-longitudinal al campo magnético externo. Con el conjunto tan magnetizado, la aplicación de un pulso de baja energía a la frecuencia de precesión de los protones, provocará una absorción de parte de toda esta energía por la muestra, haciendo inclinar los momentos magnéticos fuera de la alineación de $\vec{B_0}$ [20]. A razón de ello, habrá que

introducir la radiofrecuencia (RF), pulsos cortos de ondas electromagnéticas, que sincronizados con la frecuencia de Larmor, harán variar el estado atrás definido, inclinando el vector de magnetización sobre el plano transversal, reduciendo el longitudinal. Esto es el fenómeno de la resonancia magnética.

Ante la aplicación de un pulso de RF, se genera un segundo campo $\vec{B_1}$ cuya frecuencia coincide con la frecuencia de Larmor de los espines. La precesión de $\vec{M}$, que previamente se realizaba en el eje z (campo $\vec{B_0}$), ahora va a tener lugar alrededor del vector $\vec{B_1}$ [22]. El ángulo presente entre el nuevo eje de $\vec{M}$, y el eje de $\vec{B_0}$ (eje z), se define como ángulo de inclinación, y en función de la secuencia empleada toma unos valores u otros.

Tras el cese del pulso de excitación, el vector de magnetización neta tiende a volver a su posición de equilibrio original de forma progresiva a raíz de los dos siguientes procesos:

- **Relajación longitudinal**: recuperación gradual de la magnetización longitudinal por el proceso de intercambio de energía entre el sistema de espines y su entorno [22]. Por ello, también se le conoce como relajación espín-red.
- **Relajación transversal**: pérdida progresiva de la magnetización transversal, resultado de las distintas transiciones cuánticas que causan una transferencia de energía entre espines (desfase de los protones), y que no afectan a la energía total del sistema [22]. Por ello, también se le conoce como relajación espín-espín.

Al tiempo de relajación longitudinal se le denomina T_1, mientras que el tiempo de relajación transversal se conoce como T_2. A causa de la inhomogeneidad que presenta el campo magnético en distintas regiones de la muestra, el vector de precesión va a avanzar a diferentes velocidades en diferentes secciones de esta, produciéndose una aceleración y un consecuente valor de T_2 real, distinto y más rápido que el valor de T_2 teórico que generalmente manipulamos [22]. Este tiempo de relajación real se define como T_2^* y se determina por:

$$T_2^* = \frac{1}{\frac{1}{T_2} + \gamma \Delta \vec{B_0}} \tag{2.4}$$

$\Delta \vec{B_0}$ representa la variación de campo magnético $\vec{B_0}$ sobre la sección escaneada.

En la Figura 2.2 podemos observar la caída de T2 (relajación transversal) y la recuperación de T1 (relajación longitudinal). Debemos destacar que la relación $T_1 > T_2$ se cumple bajo cualquier circunstancia, y que no todo la magnetización perdida en el eje transversal es la recuperada en el longitudinal, si no hablaríamos de unos mismos tiempos de T1 y T2. Además, conviene recordar que los tiempos de relajación T_1 y T_2 son específicos y característicos de cada tejido.

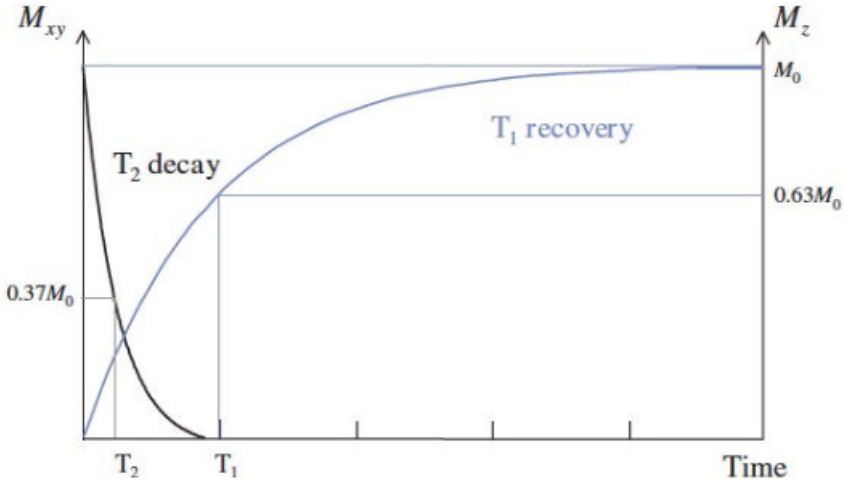


Figura 2.2: Decaimiento de T2 y recuperación de T1 [23].

2.1.3 Formación de la imagen

Una vez expuesto en el anterior apartado cómo se genera la señal de RM, en esta sección se exponen los pasos para que estas señales terminen otorgando imágenes de adquisiciones 2D.

En un primer lugar, se debe llevar a cabo la localización espacial de las señales de RM. Para lograrlo, empezamos excitando un corte específico del conjunto, en cuestión. Esto se consigue con la aplicación de un gradiente de campo magnético y un pulso de RF selectivo. Aplicando un gradiente en alguna dirección del espacio (por ejemplo, el eje z), vamos a hacer que el campo magnético varíe de manera lineal, provocando que las frecuencias de Larmor de los espines cambien dependiendo de su posición a lo largo de este gradiente [22]. En esta situación, se aplica el pulso de RF, que solo excitará aquellos espines cuya frecuencia coincida con la frecuencia de resonancia del pulso de RF aplicado. Como además, esta frecuencia de los espines es dependiente de la posición a causa del gradiente, habremos excitado solo una capa específica: nuestro corte.

Este es el principio de la excitación selectiva.

Tras esto, requerimos de la codificación de las señales inducidas para cada una de las localizaciones espaciales del corte excitado, que se logrará con gradientes adicionales, los llamados gradientes de codificación de las imágenes [4]. De no ser así, la señal de RM inducida se correspondería con la agregación de la señal generada por todos los espines, haciéndonos imposible distinguir entre las señales inducidas para cada localización. Los gradientes de codificación pueden ser de frecuencia o de fase. Si, por ejemplo, queremos obtener un corte axial, estos gradientes se van a aplicar en las direcciones x e y.

El siguiente paso es la adquisición de las muestras en el espacio K. Este espacio es una plataforma abstracta donde se adquieren, posicionan y transforman las señales digitalizadas [22]. La transformada de Fourier 2D conecta el espacio de la imagen con la señal medida en el espacio K. Para obtener la imagen final, se aplica la transformada inversa de Fourier, que transforma los datos adquiridos en este espacio de vuelta al dominio espacial, obteniendo así la reconstrucción final [4].

2.1.4 Secuencias típicas

Las IRM se obtienen a partir de secuencias de pulsos, orden específico en el que se aplican el conjunto de pulsos de RF y los gradientes de campo magnético [24]. Hay numerosas secuencias de pulso disponibles, y cada una de ellas genera una imagen con unas características propias.

Existen además unos parámetros de adquisición fundamentales, que van a ser configurados por el operador del equipo de RM para la creación de la secuencia de pulsos pertinente. Estos son:

- **Tiempo de repetición (TR)**: periodo entre pulsos de excitación de RF.
- **Tiempo de eco (TE)**: tiempo entre el pulso de excitación y el tiempo de amplitud máximo de una señal eco resultante.
- **Tiempo de inversión (TI)**: se define como el tiempo entre un pulso inicial de RF de 180^{o} y un pulso de RF posterior. Es un recurso habitual reorientar los espines en dirección

contraria a la de equilibrio, para borrar algunas de las señales de los tejidos que nos están molestando.

- **Ángulo de inclinación (α)**: representa el ángulo de giro de $\vec{M}$, con respecto al eje z (dirección de $\vec{B_0}$), tras la aplicación de $\vec{B_1}$. Este parámetro fue introducido en el apartado de bases físicas.

A continuación, se detallan las secuencias de pulso más comunes, que también son aplicadas para la obtención de la base de datos de este trabajo:

- **Secuencia eco de espín** (**SE**, de las siglas en inglés *Spin Echo*): es la secuencia más común. Utiliza dos pulsos de RF: un primero de excitación de 90º, que gira los espines hacia el plano transversal, y después, un pulso de 180º para volver a alinear entre sí los espines y formar una señal eco [24]. Este pulso de re-enfoque, va a introducir una inversión en fase, devolviéndole la coherencia tras el paso de un determinado tiempo desde su aplicación (TE/2). El resultado es una señal de eco, que será la que finalmente se recoja. En la Figura 2.3, observamos los estados de magnetización para un TR.

- **Secuencias de eco de gradiente** (**GRE**, de las siglas en inglés *Gradient echo*): son un concepto diferente, en tanto que usan un par de gradientes de campo magnético bipolares para la producción de un eco después del pulso de excitación, en lugar del re-enfoque de 180º realizado para otras secuencias [24]. Debido a la evitación de este pulso, podemos lograr TRs más cortos, y tener unos tiempos de adquisición para esta secuencia mucho menores que para otras.

- **Secuencia de recuperación por inversión** (**IR**, de las siglas en inglés *Inversion recovery*): es una secuencia eco de espín a la que se le ha añadido un pulso de inversión de 180º previo al pulso de excitación de 90º de la secuencia SE [24]. Se puede utilizar también con GRE. Estas secuencias son típicamente utilizadas para la imagen T1w, y dos de los ejemplos más comunes son la secuencia de recuperación por inversión de tiempo corto (STIR, de sus siglas en inglés *Short tau inversion recovery*), y la secuencia de recuperación por inversión con atenuación de fluido (FLAIR, de sus siglas en inglés *Fluid attenuated inversion recovery*).

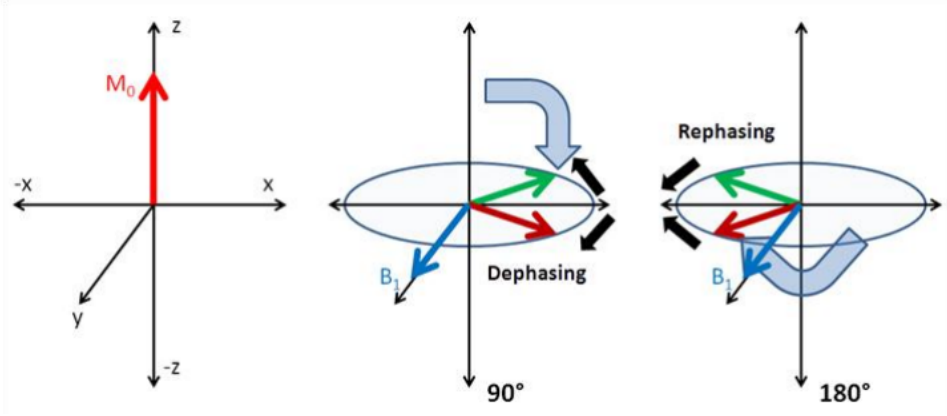


Figura 2.3: Estados de magnetización en una secuencia eco de espín [21].

2.1.5 Ecuaciones teóricas de las secuencias de pulsos

Las secuencias de pulsos presentan una solución analítica específica que, en casos simples, podemos utilizar para sintetizar imágenes ponderadas por RM [4]. Estas ecuaciones están definidas por los parámetros de adquisición anteriormente definidos (TR, TE, TI y α), y por los mapas paramétricos T1, T2 y PD. En el trabajo realizado se sintetiza la imagen post-T1w y la T2w a partir de estos mapas, que junto con los parámetros de adquisición propios de estas imágenes (véase el Anexo A), nos permiten utilizar las ecuaciones definidas:

$$m_{\text{IR-GRE}}(x) = PD(x)\frac{\left(1 - 2e^{-\frac{TI}{T1(x)}} + e^{-\frac{TR}{T1(x)}}\right)}{1 + \cos(\alpha)e^{-\frac{TR}{T1(x)}}}\sin(\alpha)e^{-\frac{TE}{T2(x)}} \tag{2.5}$$

$$m_{\text{SE}}(x) = PD(x)\left[1 - 2e^{-\frac{TR-\frac{TE}{2}}{T1(x)}} + e^{-\frac{TR}{T1(x)}}\right]e^{-\frac{TE}{T2(x)}}, \tag{2.6}$$

siendo x la localización de un voxel del total de la imagen.

La primera ecuación se corresponde con la secuencia IR-GRE, utilizada para la post-T1w, y la segunda se corresponde con la SE, ecuación de pulsos aplicada para la T2w.

2.2 Agentes de contraste

La alta resolución espacial y la capacidad de definición y distinción de los tejidos blandos son las principales ventajas de la MRI. No obstante, habrá ocasiones en las que el contraste observado no será suficiente, y será con la aplicación de agentes de contraste (CAs, de las siglas en inglés *contrast agents*) como conseguiremos mejorarlo.

Los iones metálicos con uno o más espines desapareados son paramagnéticos, y por tanto, poseen un momento magnético permanente, como hemos visto en las bases físicas de la RM. El gadolinio (Ga^{3+}) y el manganeso (Mn^{2+}) son ejemplos de iones paramagnéticos que se utilizan como CAs en RM. Una de las familias de CAs más utilizadas, y también aplicada en este trabajo, es la familia de agentes de contraste basados en gadolinio, GBCA.

Estos iones mencionados poseen espines no apareados en sus orbitales externos, permitiendo ser atraídos por campos magnéticos. De este modo, iones con estas características van a potenciar las fluctuaciones en la interacción magnética en el proceso de recuperación, reduciendo así los tiempos de relajación T1 y T2 en el tejido objetivo [25]. El T1 siempre es un valor considerablemente mayor que el T2, por lo que en la clínica, el mayor impacto que tienen los CAs en dosis bajas consiste en el acortamiento de T1 [4]. Por este motivo, los tejidos que acumulan estos agentes aparecen hiperintensos en las imágenes T1w, siendo la post-T1w de gran valor diagnóstico.

El procedimiento estándar utilizado en la clínica parte de la adquisición de una imagen T1w pre-contraste, a la que se le administra un GBCA mediante inyección intravenosa, para finalmente, tras una espera de 5-10 minutos, realizar la adquisición de una imagen ponderada en T1 post-contraste (post-T1w), que será comparada con la T1w de referencia de forma subjetiva y visual, por el especialista [4].

La barrera hematoencefálica (BBB, de sus siglas en inglés *blood-brain barrier*) es la interfaz física entre el sistema nervioso central (SNC), y la circulación sistémica, y va a actuar de regulador de la homeostasis del SNC conservando la función normal del mismo. Durante la progresión de una patología tumoral cerebral, la BBB experimenta alteraciones, dando lugar a lo

que se denomina barrera hematoencefálica tumoral (BTB, de sus siglas en inglés *blood-tumor barrier*), un estado patológico de la BBB [26]. La BTB se caracteriza por ser significativamente más permeable que la BBB, permitiendo un mayor paso de sustancias y componentes entre el torrente sanguíneo y el parénquima cerebral. Entre estas sustancias, también se encuentran los CAs. La inyección de agente en sangre se va a traducir en una fuga del mismo por el parénquima en las zonas con la BBB dañada, y en una consecuente visualización de realce en esas zonas en las imágenes post-contraste adquiridas (véase la Figura 2.4).

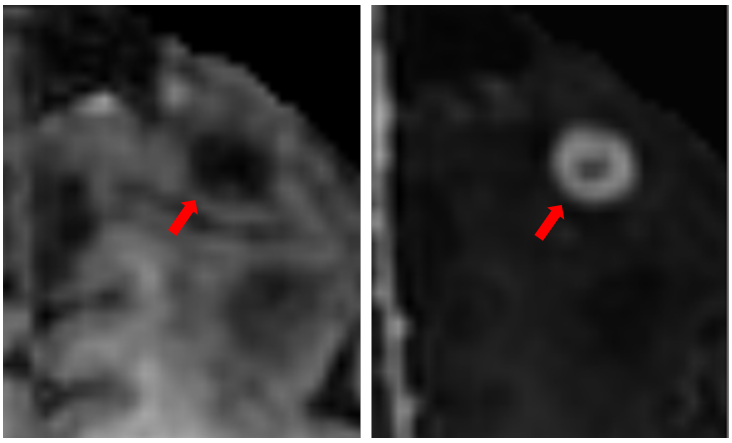


Figura 2.4: T1w (izquierda) y post-T1w (derecha), mostrando realce por GBCA de una misma región cerebral en un sujeto de nuestra base de datos.

Cerrando este apartado, debemos realizar una reflexión sobre las **limitaciones** y dificultades asociadas a la aplicación de CAs. Estas han sido fundamentales para la justificación del trabajo, y razón de ello su exposición en la sección de motivaciones del mismo. A continuación se presentan:

- Reacciones adversas de los pacientes [9].
- Situaciones aparatosas e incómodas para el paciente, durante la inyección intravenosa.
- Riesgo de retención (almacenamiento) de GBCA en los tejidos del cerebro [10].
- Impacto medioambiental: contaminación de aguas superficiales, aguas potables, sedimentos, e incluso de organismos vivos [11].
- Costes: los CAs poseen un coste elevado, que sumado al personal requerido especializado en la materia, encarece sustancialmente la adquisición de imágenes post-contraste.

2.3 Aprendizaje profundo

Actualmente, los sistemas inteligentes que ofrecen capacidades de inteligencia artificial suelen estar basados en el aprendizaje automático. El aprendizaje automático o *machine learning* se describe como la capacidad que poseen los sistemas para aprender a partir de datos específicos de un problema (de entrenamiento), con lo que poder automatizar los procesos de construcción de modelos, y resolver las tareas asociadas [12]. Este estudio se basará en el uso del aprendizaje profundo (DL, de sus siglas en inglés *deep learning*), un método del aprendizaje automático que tiene como fundamento las redes neuronales artificiales.

2.3.1 Redes neuronales artificiales

La unidad funcional del DL son las redes neuronales artificiales (ANNs, *Artificial Neural Networks*), resultado de la admiración por el funcionamiento del cerebro humano y su capacidad para realizar tareas complejas en un corto espacio de tiempo y con una alta eficiencia [27]. Algunas de las características más importantes de las neuronas son: la alta interconexión existente entre ellas, lo que les permite recibir estímulos de eventos recién ocurridos, pero también cientos de señales eléctricas con información aprendida, el trabajo en paralelo, y la organización en capas consecutivas [27]. La asociación de las neuronas se conoce como redes neuronales biológicas (BNNs, *Biological Neural Networks*),y son la inspiración de las ANNs. En la Figura 2.5, observamos una BNN y su organización por capas consecutivas.

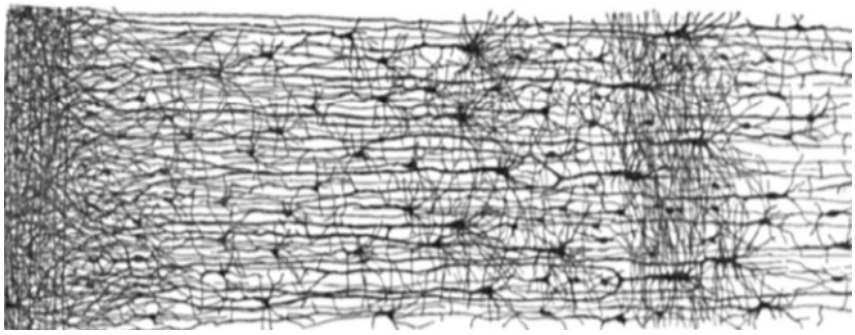


Figura 2.5: Múltiples capas de una red neuronal biológica de la corteza cerebral [27].

Las ANNs están compuestas por una capa de entrada, que se encargará de recibir los datos en forma de valores numéricos, varias capas ocultas que realizan los cálculos, y una capa de salida que va a realizar la predicción, de la índole que sea [4].

Se implementan desarrollando un algoritmo de aprendizaje computacional que no necesita programar sus reglas, sino que es capaz de construir sus reglas de comportamiento a través de la denominada "experiencia". Estos algoritmos son sistemas de computación masivamente paralelos, compuestos por una gran cantidad de unidades de procesamiento básicas (las neuronas), interconectadas entre sí y con adquisición de experiencia por su entorno [27]. El factor clave de este aprendizaje vendrá mediado por los pesos sinápticos de la red, encargados de modificar la información recibida emulando el fenómeno sináptico, es decir, aportarán una mayor o menor fuerza de conexión a la relación entre neuronas, modificando así la influencia que una tiene sobre otra, en función de su valor [27]. La tarea del algoritmo será la de ajustar estos pesos de manera secuencial y controlada para alcanzar el objetivo señalado.

En la Figura 2.6 se establecen los principales elementos de los modelos de ANNs:

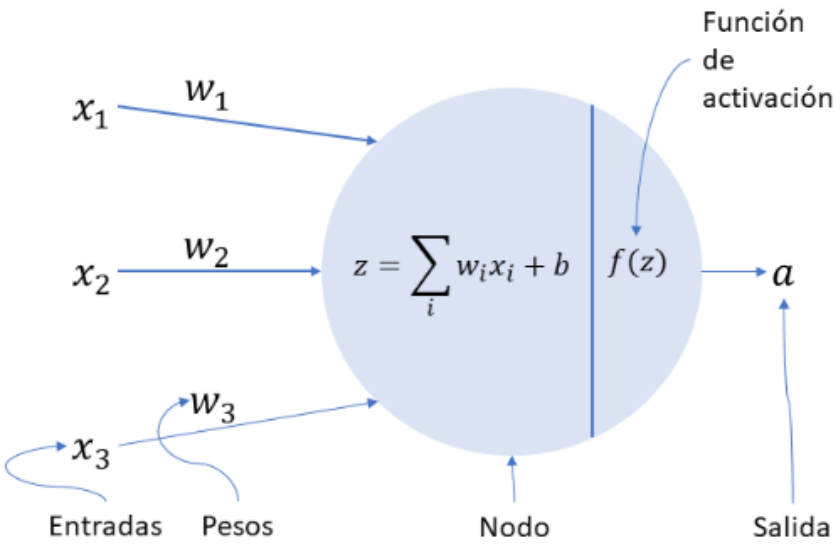


Figura 2.6: Elementos de una ANN [28].

En primer lugar las entradas (x_1, ..., x_n), representan los estímulos que la neurona recibe, a través de las conexiones interneuronales que presenta. Cada uno de estas entradas va a ser multiplicada por su correspondiente peso (w), el cual ajusta la relevancia de la información proveniente. Por último se suma el sesgo o *bias* (b), que actuará como umbral. El resultado de esta operación es z (véase la ecuación (2.7)), que será posteriormente introducido en una función de activación, f, que decidirá si la neurona se activa, o no. Existen variadas funciones

de activación, pero definiremos la ReLU (ecuación (2.8)), la preferida para este proyecto.

$$z = \sum_{i=1}^{n} x_i w_i + b \tag{2.7}$$

$$Y = \begin{cases} 0, & \text{si} \quad z \leq 0 \\ z, & \text{si} \quad z > 0 \end{cases} \tag{2.8}$$

Como podemos deducir de la ecuación, la activación solo se producirá cuando z sea superior a 0, devolviendo el mismo valor de salida que dicho z, en estos casos. Es tras esta evaluación, cuando se produce la salida de la red (Y).

2.3.2 Tipos de aprendizaje

Podríamos definir el entrenamiento de una red neuronal como el proceso iterativo de optimización que tiene por finalidad ajustar el modelo con el que realizar las predicciones, en base a unos datos determinados de entrada. Los tipos de aprendizaje en función de las características del proceso de entrenamiento son extrapolables a experimentos realizados más allá del aprendizaje profundo, en cualquiera de sus métodos y campos. Los desarrollamos a continuación:

- **Aprendizaje supervisado**: utiliza datos de entrenamiento etiquetados para entrenar el modelo. En otras palabras, las predicciones que el modelo genera durante el entrenamiento son comparadas con las etiquetas conocidas para ajustar los parámetros del modelo y mejorar su rendimiento. Con ello, se pretende crear modelos capaces de generalizar, de modo que puedan realizar predicciones precisas cuando se les presentan nuevos datos.

- **Aprendizaje no supervisado**: el entrenamiento se sucede a partir de datos sin etiquetar. La idea detrás de este aprendizaje es aprender patrones de los mismos datos de entrenamiento, y a partir de ellos agruparlos en distintas categorías.

- **Aprendizaje semi supervisado**: se sitúa entre el aprendizaje supervisado y el no supervisado. En este caso, solo un conjunto mínimo de datos estará etiquetado, y será el propio algoritmo el que vaya realizando inherencias y generando automáticamente nuevas etiquetas, gracias a la información inicial otorgada. Resulta de gran utilidad cuando la obtención de datos etiquetados viene asociada con unos costes elevados, reduciendo así el impacto.

- **Aprendizaje autosupervisado**: variación del aprendizaje no supervisado, que permite la obtención de conocimiento de los datos de entrada, evitando la necesidad de etiquetado explícito. Suele estar asociado a redes que incorporan conceptos y contenidos físicos durante el proceso de entrenamiento.

- **Aprendizaje por refuerzo**: es un aprendizaje por prueba y error. Un agente aprende a realizar una tarea dentro de un ciclo de retroalimentación, y se repite hasta que el resultado obtenido entra dentro de unos márgenes aceptables (los que se marquen). Ante una tarea realizada correctamente, el agente recibe un refuerzo positivo. El aprendizaje de los perros es un claro ejemplo de ello.

Los dos aprendizajes utilizados en este Trabajo de Fin de Grado han sido el supervisado y el autosupervisado.

2.3.3 Redes neuronales convolucionales

Las redes neuronales convolucionales (CNNs, de sus siglas en inglés *Convolutional Neural Networks*), son una forma especializada de redes neuronales que utiliza datos de entrada con una estructura espacial [27].

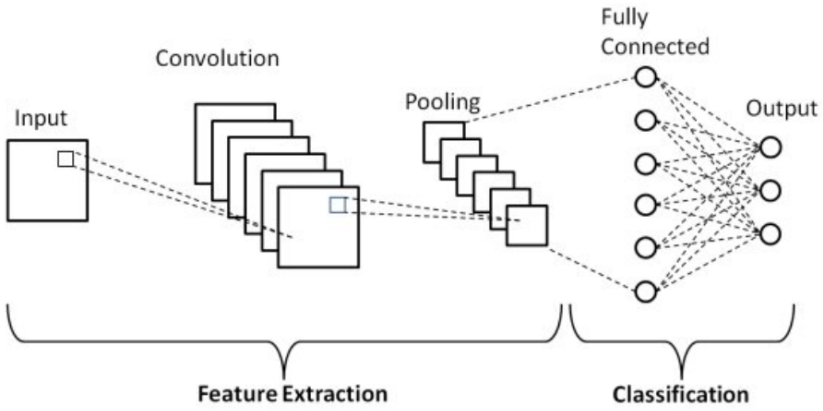


Figura 2.7: Arquitectura básica de una CNN [29].

La estructura de red de la Figura 2.7 muestra las capas más comunes de las CNNs. Se detallan a continuación estas capas:

- **Capas convolucionales**: la convolución es una operación matemática que involucra un filtro o *kernel*, que va a desplazarse sobre la entrada para generar una salida conocida como mapa de características. El *kernel* es una matriz cuadrada que va a utilizarse como operador para la entrada, obteniendo a la salida el resultado de su aplicación espacial en ella. Algunos de los hiperparámetros más importantes en una capa convolucional son el *stride* y el *padding*. El *stride* define el paso que da el *kernel* sobre la entrada. Para un *stride* de 1, el filtro se va a desplazar píxel por píxel, realizando la convolución. Por último, surge el *padding*, que es una forma de control de la dimensión. La propia operación de convolución reduce la salida tal que:

$$\text{Salida} = \left\lfloor \frac{\textit{tamaño entrada - tamaño kernel}}{\textit{stride}} \right\rfloor + 1 \tag{2.9}$$

  El padding tendrá, de forma general, dos opciones: *valid padding*, que habilita la reducción de la dimensión a partir de la ecuación 2.9, o *same padding*, que va a conservar la dimensionalidad tras la convolución, mediante la adición de 0's.

  Es común tras aplicar la convolución, introducir una función de activación, que introduzca la no linealidad en el modelo. La misma capa convolucional de *TensorFlow* permite añadir un argumento de activación, especificando el tipo de función ($ReLU$ en el trabajo).

- **Capas de *pooling***: reemplaza la salida de la operación de convolución en una localización determinada, con estadísticas de los vecinos o salidas cercanas. Van a reducir la dimensionalidad de la muestra y también vendrá definido por el tamaño del *kernel*. Popularmente existen dos operaciones de *pooling*, el *max pooling* y el *average pooling*. El primero escogerá el mayor de los números que se encuentren en el área del *kernel* (véase la Figura 2.8), mientras que el segundo realizará la media de todos los valores de ese área del filtro. También se fija un *stride* con el mismo funcionamiento que las capas convolucionales.

- **Capas densas**: cada neurona está conectada a todas las neuronas de la capa anterior y de la capa siguiente (véase Figura 2.7, de ahí que también sean conocidas como capas *fully-connected*. Presentan un papel generalmente clasificador, presentándose en las últimas capas de las redes CNN, generalmente. También van asociadas a una función de activación,

y en este caso, también se requiere del número de neuronas de salida, es decir, el número de posibilidades de clasificación que hay.

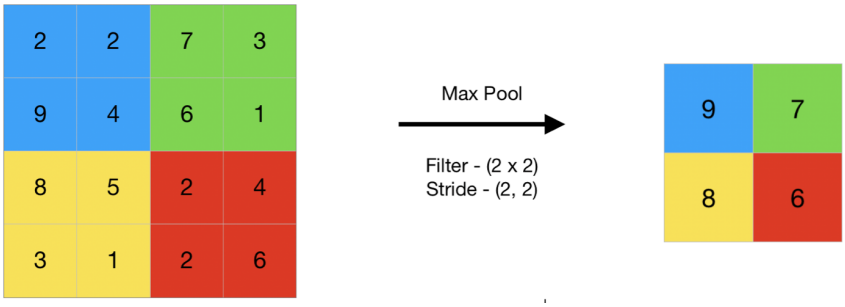


Figura 2.8: *Max pooling* con filtro y *stride* de tamaño 2 [30].

Las capas convolucionales y *pooling* han sido ampliamente utilizadas en el Trabajo de Fin de Grado, no así las capas densas, cuyo propósito no está relacionado con el enfoque de síntesis de imagen que se ha llevado a cabo a partir de las técnicas de DL aplicadas.

2.3.4 Arquitectura U-Net

U-Net es una arquitectura de red neuronal totalmente convolucional, originalmente orientada a la tarea de segmentación de imágenes. Fue introducida en el año 2015 en el artículo *"U-Net:Convolutional Networks for Biomedical Image Segmentation"* [31] y está compuesta por dos partes principales: el *encoder* (contracción), y el *decoder* (expansión) [32]. La parte de contracción será responsable de identificar características relevantes de las imágenes de entrada, aumentando mediante convoluciones la profundidad (número de canales) de los mapas de características, y reduciendo el ancho y alto. Tras esto, se produce la etapa expansiva, donde la red buscará localizar las características, manteniendo la resolución espacial en todo momento. Las conexiones de salto desde la ruta de contracción servirán para conservar la información espacial que se pierde en esta misma ruta, lo que permite que las capas de decodificación localicen las características con mayor precisión [32]. Todo este proceso podemos observarlo en la Figura 2.9.

La segmentación es el principal campo de especialización de estos tipos de red, sin embargo, estudios recientes han aplicado esta arquitectura (y en simbiosis con otras) al ámbito de la síntesis de imagen [33, 34], y en este trabajo buscaremos su aplicación y rendimiento en este campo también.

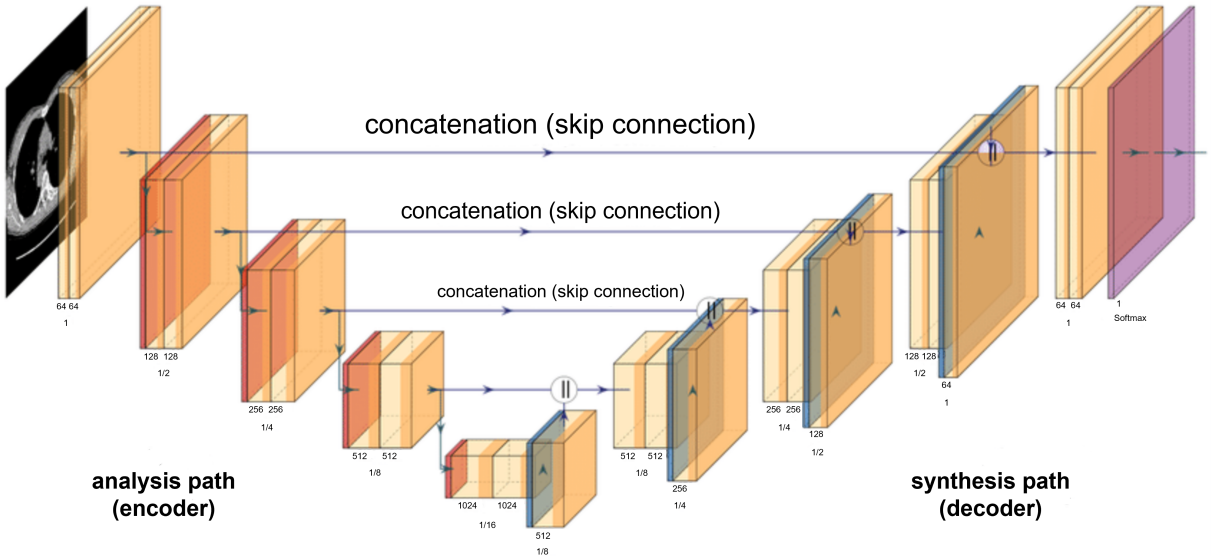


Figura 2.9: Esquema volumétrico de una red 3D U-Net [35].

2.3.5 Transformadores de visión

Los transformadores de visión o ViT, de sus siglas en inglés *Vision Transformers*, son una tecnología de redes neuronales basadas en la auto-atención. Surgen de los transformadores originales, cuyo mecanismo se basaba en la atención para establecer dependencias globales entre la entrada y la salida, y que tenían su aplicación en el procesamiento de lenguaje natural, y en todo lo que envolviera a textos, en inteligencia artificial. Este nuevo paradigma de ViT, con sus modificaciones, introduce el análisis de imagen en el aprendizaje profundo.

Esta arquitectura va a empezar dividiendo la imagen (o entrada), en pequeños parches (similar a las palabras en el procesamiento de lenguaje natural), para después ser linealmente vectorizados. Es decir, toda la información contenida en el parche se va a convertir en un vector lineal. Tras esto, se añaden los *positional embeddings* al vector, que son valores que van a indicar la posición relativa del parche dentro de la imagen. Esto es una característica fundamental de los ViT porque los transformadores originales no consideran la estructura espacial de la imagen, y con estos *embeddings*, la conservamos. Desde este momento, la secuencia de vectores resultante se va a aplicar a un codificador transformador estándar (introducción a una pila de capas de un transformador). La principal característica de este paso son las capas de auto-atención, pilar de los transformadores.

La auto-atención es el mecanismo que permite al modelo considerar las relaciones entre todos los parches, de forma simultánea. Nos ayudará a capturar dependencias globales en la imagen, permitiéndonos relacionar regiones de cualquier parte de la imagen, y perder la dependencia obligatoria de la proximidad, propia de las CNNs.

Aunque los ViT, dentro de su aplicación en imágenes, están más enfocados en la clasificación de imágenes, en este Trabajo de Fin de Grado se propone una red que combina la CNN de arquitectura U-Net con ellos, para la síntesis de imágenes de RM. Tenemos constancia de la aplicación en este campo gracias a distintos trabajos [36, 37].

Capítulo 3

Métodos

En este capítulo se exponen los pasos realizados para la consecución de los objetivos, desde los datos de partida iniciales, hasta la obtención de las imágenes sintéticas para los experimentos y técnicas aplicadas, y sus métricas de calidad.

3.1 Preprocesamiento de la base de datos

Como hemos explicado en varias ocasiones durante esta memoria, partimos de una base de datos de imágenes ponderadas T1w, T2w, T2w-FLAIR y post-T1w, y de unos mapas paramétricos T1, T2 y DP.

En relación al preprocesamiento de los datos, detallamos los pasos dados a continuación:

1. Orientación de todas las modalidades de imagen siguiendo el atlas MNI152 [38].
2. Aplicación de la herramienta HDBET para la retirada de tejido no cerebral de las imágenes [39].
3. Registro de las imágenes del cerebro obtenidas con las imágenes T1w aplicando la herramienta FLIRT del *software* FSL [40]. Con este paso, efectuamos que todas las imágenes estén alineadas en un mismo espacio de referencia.
4. Normalización de las imágenes ponderadas preprocesadas (T1w, T2w, T2w-FLAIR, post-T1w) dividiendo por su intensidad promedio, excluyendo los valores nulos de la imagen previo a la normalización. En el apartado cuantitativo, la normalización de los mapas paramétricos consistió en una conversión de escala de milisegundos a segundos para el T1

y T2 (división por 1000), y de una división por 100 para el DP, con el objetivo de que sus valores oscilen en una escala entre 0 y 1.

5. Extracción de los primeros 40 y de los últimos 30 cortes de cada una de los volúmenes 3D de los pacientes. Con el fin de reducir carga computacional, se prescinde de 70 cortes de fondo.

Por otro lado, nuestra base de datos fue introducida en la herramienta de segmentación de aprendizaje profundo HD-GLIO [41], con la que se obtuvieron segmentaciones reales (máscaras) de las regiones con realce y edema tumoral para el cálculo de métricas de calidad. HD-GLIO utiliza imágenes pre-contraste y post-contraste como entradas, y es importante destacar que fue entrenada con anotaciones realizadas por expertos [4].

3.2 Construcción de entradas y salidas del modelo

Como hemos desarrollado, el conjunto de datos utilizado proviene de 14 pacientes. Cada uno de ellos incluye las modalidades de imagen anteriormente mencionadas. De esta forma, para cada modalidad disponemos de un volumen tridimensional que sigue la forma: número de cortes axiales (profundidad) x 256 x 256. El ancho y largo de cada corte, y por tanto, del volumen 3D es de 256 píxeles por 256 píxeles.

Nuestros datos de entrada son los mapas paramétricos T1, T2 y DP. El enfoque que se ha seguido para la construcción del conjunto de entrenamiento (x_{train}) y del conjunto de validación (x_{valid}), ha sido combinar todos los cortes de cada paciente en una única matriz de datos. Esto se ha logrado tomando los cortes axiales de todos los pacientes y concatenándolos, por mapa. Es decir, el resultado de la primera dimensión de nuestro conjunto de entrenamiento es del tamaño de la suma de todos los cortes de los pacientes de entrenamiento (ídem para la validación) , concatenando cada volumen después del anterior, mientras que el ancho y alto de la imagen se conserva. El último paso dado ha sido la concatenación en una última dimensión de los 3 mapas, por corte (véase figura 3.1). Sabemos que T1, T2 y DP están relacionados corte a corte en nuestra base de datos, por tanto, cada uno de ellos (el corte exacto de cada uno de los 3 mapas) se juntó en una última dimensión, que pasó a tener un valor de 3 (canales). De esta forma, se obtienen tantos *stacks* de 256 x 256 x 3, como cortes hayamos obtenido en la primera dimensión.

Así, el tamaño de x_{train} y x_{valid} viene dado por la forma:

$$x_{\text{train}} = \text{número cortes } \textit{train} \text{ total} \times 256 \times 256 \times 3$$
$$x_{\text{valid}} = \text{número cortes } \textit{valid} \text{ total} \times 256 \times 256 \times 3$$

Por otro lado, nuestro conjunto de test es un *stack* de las mismas características, y la primera dimensión (número de *stacks*) el resultado de la profundidad de la imagen del paciente test.

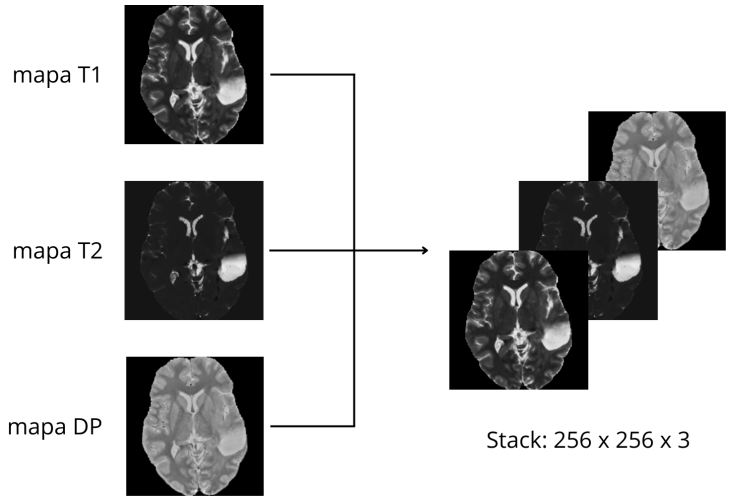


Figura 3.1: Creación de *stack* de entrada.

Se ha procedido de manera paralela con el conjunto de etiquetas de entrenamiento (y_{train}) y validación (y_{valid}). En este caso, solo encontramos un canal (post-T1w adquirida), por lo que la única concatenación a realizar es la de la suma de los cortes de todos los pacientes en la primera dimensión, la cual debe ser respetada entre los datos de entrada y las etiquetas para el conjunto de entrenamiento y de validación.

Toda esta disposición planteada ha permitido transformar los datos originales, que estaban organizados por paciente, en un solo conjunto grande, adecuado para el entrenamiento de la red neuronal. Además, para una mayor comodidad, y una reducción de los tiempos de entrenamiento, los conjuntos definidos fueron almacenados en ficheros *.npz* para cada paciente. De esta forma, no requerimos de la realización de parte del preprocesado ni de la reconstrucción en cada prueba que se haga, sino que ya tenemos estos ficheros almacenados con todos los pasos dados.

3.2.1 Imagen completa

El entrenamiento por imagen completa consiste en la utilización de los datos almacenados, con la resolución del corte completo. Se detalla un ejemplo, a continuación para el primer paciente:

$$x_{\text{train}} = 3182 \times 256 \times 256 \times 3$$
$$y_{\text{train}} = 3182 \times 256 \times 256 \times 1$$
$$x_{\text{valid}} = 3182 \times 256 \times 256 \times 3$$
$$y_{\text{valid}} = 3182 \times 256 \times 256 \times 1$$

La entrada de la red para la imagen completa es:

$$\textit{Entrada} = \textit{batch size} \times 256 \times 256 \times 3$$
$$\textit{Salida} = \textit{batch size} \times 256 \times 256 \times 1$$

Nuestro conjunto viene dado por el tamaño de lote, comúnmente conocido como *batch size*. Este hiperparámetro se refiere a la cantidad de ejemplos de entrenamiento que el modelo procesa antes de actualizar sus parámetros internos (pesos y sesgos), durante cada paso de optimización. En este trabajo se ha fijado un *batch size* de 64, por lo que nuestro entrenamiento, para este primer sujeto, cuenta con 50 pasos (*steps*) por época (el último incompleto). Los pasos se corresponden con el número de actualización de parámetros que se realiza.

3.2.2 Basado en parches

Con el propósito del trabajo de buscar el detalle en regiones específicas, se pasó a la implementación de experimentos con entradas por parches. La construcción de los mismos se realizó a partir de los *stacks* generados, visualizable en la Figura 3.2.

Las dimensiones del parche, observable en el margen inferior de la Figura 3.2, nos indican que el tamaño de parche elegido ha sido de 64, mientras que la última dimensión se conserva por la concatenación de los tres mapas paramétricos. El tamaño elegido se debe a que es un divisor exacto de 256, y para la evolución explicada en el siguiente párrafo, es el más conveniente.

En un primer momento, se trabajó con parches exactos para el tamaño total de la imagen. Es decir, para cada *stack* de tamaño 256 x 256 x 3, generábamos 16 parches 64 x 64 x 3 (véase Figura 3.3). Para un mismo *batch size*, el número de pasos aumentaba considerablemente. Con

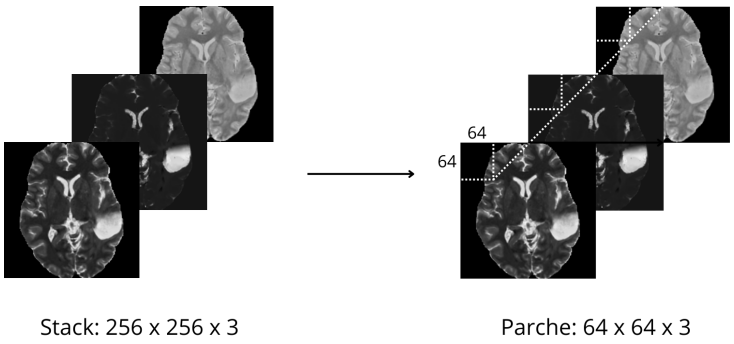


Figura 3.2: Planteamiento de la construcción de parches.

un ejemplo práctico, para 3182 *stacks* originales, se obtienen 3182 x 16 parches (50912 parches). Estos son introducidos a la red, cuyo formato es ahora: $Entrada = batch\ size \times 64 \times 64 \times 3$. Donde antes encontrábamos 50 *steps*, ahora tenemos 796.

La razón del tamaño 64 se debe a que es un divisor exacto de 256, y por tanto, nos permitía la descomposición perfecta de las imágenes originales. Sin embargo, todo este proceso traía consigo un artefacto que lastraba mucho la visualización de la imagen, el efecto "de bloques". Cuando las predicciones de cada parche del conjunto de test eran montadas una detrás de otra, se generaba este efecto de bloques, desmeritando los avances obtenidos con esta evolución por parches.

Para evitarlo, se planteó una construcción de parches con solapamiento. Esta nueva versión, permitiría suavizar los bordes al compartir los parches contiguos valores durante varios píxeles. El solapamiento establecido fue de 16, de forma que por cada *stack* ya no se generan 16 parches, sino 25 parches 64 x 64 x 3. En función de la localización del parche original pueden estar dos, tres o cuatro de los lados del parche solapados, siguiendo la Figura 3.3. Se han obtenido mejoras sustanciales en materia del efecto "de bloques", y ha sido finalmente el formato aplicado para las entradas por parches en este trabajo.

En la Figura 3.3 observamos en el margen izquierdo el planteamiento original, y en el derecho, la construcción final, con un aumento en la intensidad de las zonas solapadas para mejor comprensión. Debemos tener en cuenta que, aunque estemos viendo dos figuras bidimensionales, poseen una profundidad de 3 canales que no ha sido representada.

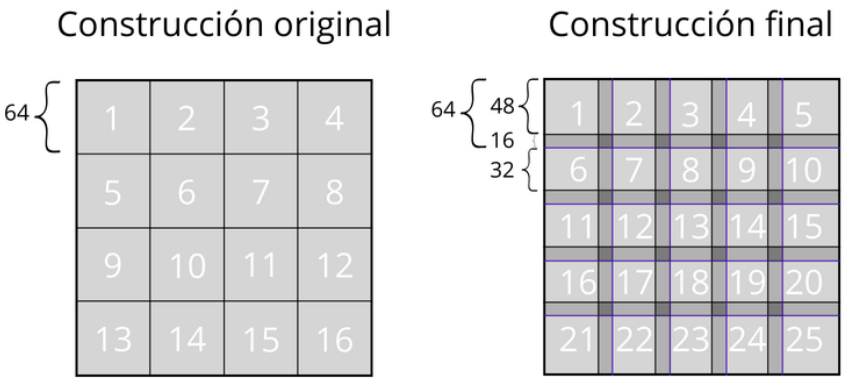


Figura 3.3: Evolución de la construcción de los parches.

3.2.3 Aumento de datos de parches con realce

Durante el proyecto, se realizó un experimento con un aumento de datos enfocado en la mejora del experimento que mejores datos estaba otorgando hasta el momento (autosupervisado por parches). Esta prueba consistió en introducir de nuevo los parches ubicados en regiones de realce de las máscaras de los pacientes. A partir de las máscaras de realce y edema tumoral (con valores nulos fuera de estas regiones), es posible extraer los parches cuya ubicación espacial coincida con área de realce. Para lograrlo, se aplicó un filtro que desechaba, del total de parches, aquellos cuyo valor en en la región de máscara asociado a este, fuera nulo. Esto generó un total aproximado de 5000 parches de regiones con realce (o edema).

Para aumentar más los datos, se introdujeron en una función que duplicaba cada parche mediante un giro de 20º, y se pasó por esta función dos veces. En total, se generaron unos 20,000 parches, que fueron concatenados al conjunto de entrenamiento con los 80,000 originales ($\approx 3200\,stacks \times 25\,\text{parches}/stack$). Para el grupo de validación se llevó a cabo el mismo procedimiento.

3.3 Arquitecturas de redes neuronales aplicadas

Las redes neuronales utilizadas han sido dos:

- CNN con arquitectura U-Net.
- CNN con arquitectura U-Net, incorporando ViTs.

3.3.1 Arquitectura U-Net

La arquitectura planteada en este trabajo es una implementación en el marco de la síntesis de imagen de la U-Net original desarrollada por científicos de la Universidad de Freiburg para segmentación de imagen [31]. La estudiamos a través de la siguiente Figura 3.4:

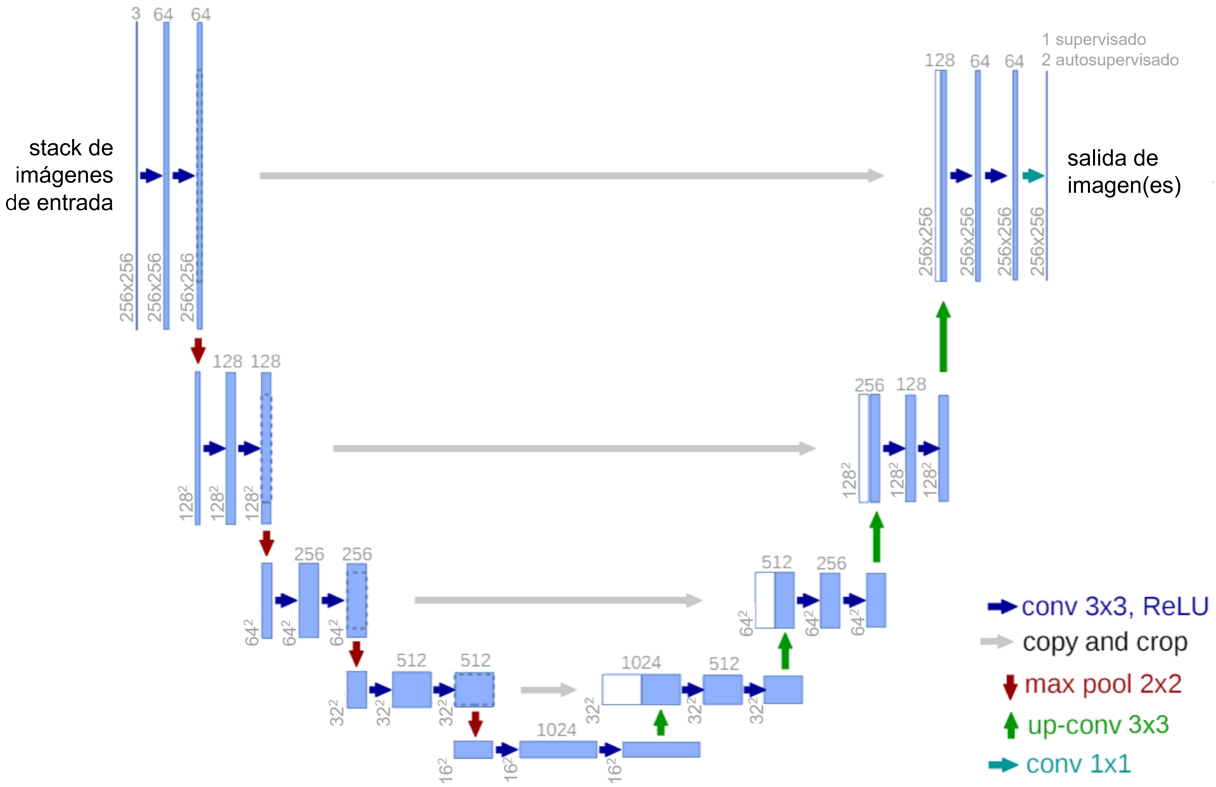


Figura 3.4: Arquitectura U-Net utilizada en el trabajo para experimentos de imagen completa [31].

Los cambios aplicados en la red, que pueden ser observados en la Figura 3.4, han sido la resolución de los mapas de características en todas sus capas, pues partimos de unas imágenes diferentes, el número de canales de entrada (3), el de salidas en función del aprendizaje aplicado (desarrollado en la Sección 3.4), y lo que se define como *up-conv*, la convolución transpuesta con la que se duplica la resolución de las características, que se ha fijado en un tamaño de filtro de 3x3, mientras que el artículo original la realiza con un 2x2. Otro cambio importante introducido ha sido la aplicación de *same padding* en las convoluciones 3x3, en lugar de *valid padding*. Además, se han introducido distintas capas de *dropout* que aparecerán a medida que vayamos exponiendo nuestra U-Net.

Como podemos observar, la arquitectura consiste en un camino de contracción y en un camino de expansión, el ***encoder*** y ***decoder***, respectivamente. En el margen inferior de la imagen, observamos el llamado ***bottleneck*** o cuello de botella. Pasamos a desarrollar cada uno de estos caminos:

El *encoder* consiste en la aplicación repetida (4 veces) de un bloque (*downsample block*) cuya composición podemos definir en:

1. Aplicación de dos convoluciones 3x3, con *same padding* (conservando la dimensionalidad) y función de activación $ReLU$. El número de filtros definido para estas dos convoluciones se corresponde con el número de canales de características de la salida del bloque. Para el primer bloque, siempre se fija 64 como número de filtros de salida.

2. Operación de *max pooling* 2x2 con un *stride* de dos. Esta operación es la que va a devolver una reducción del 50 % en la resolución de la salida, en comparación con la entrada.

3. Introducción de un *dropout* del 30 %. El *dropout* va a "apagar" el 30 % de las neuronas que llegan a esta capa de forma aleatoria, para evitar el sobreajuste.

Como resultado de estos pasos, obtenemos una duplicación en el número de canales, y una reducción del 50 % en la resolución de la entrada. Para 4 bloques (repeticiones), la entrada y salida del camino de contracción de la arquitectura U-Net se corresponderá con:

$$\text{Entrada encoder} = \text{batch size} \times 256 \times 256 \times 3$$
$$\text{Salida encoder} = \text{batch size} \times 16 \times 16 \times 512$$

Estas dimensiones se establecen para los experimentos de imagen completa, que difieren en resolución de los de entrada basada en parches. Para este formato, la entrada y salida del *encoder* viene dada por:

$$\text{Entrada encoder} = \text{batch size} \times 64 \times 64 \times 3$$
$$\text{Salida encoder} = \text{batch size} \times 4 \times 4 \times 512$$

En última instancia, es fundamental destacar la construcción de los tensores f y p en la etapa de contracción. A las salida de las capas convolucionales en cada bloque, se obtendrá un tensor f que será pasado por la capa de *max pooling* y *dropout*, denominándose en este momento p (f no se ve alterado por estas dos capas). Tanto f como p serán dos tensores producidos en

cada iteración, de mismo número de canales, y distinta resolución. El tensor p llevará la información por toda la arquitectura, aplicándose en el *decoder* y cuello de botella, mientras que f se almacena hasta su concatenación en la etapa de expansión. Esta concatenación es la clave para recuperar el tamaño original en el *decoder*.

Avanzando en la arquitectura, nos encontramos con el cuello de botella, definido como un camino horizontal en el borde inferior de la Figura 3.4. Es muy similar a las capas del *encoder*, desmarcándose de él por la ausencia de las capas de *max pooling* y *dropout*. Este es el estado con la representación más comprimida de la imagen, y desde este punto no se va a reducir más la dimensión. Únicamente se aplicarán las dos capas convolucionales 3x3, duplicando de nuevo el número de parámetros hasta 1024. En este tramo trataremos de capturar las características más abstractas, previo a la fase de decodificación.

Por último, tenemos el *decoder*, nuestro camino de expansión en la arquitectura. De igual modo que para la contracción, se basa en un bloque (*upsample block*) que se itera cuatro veces. Exponemos el bloque al detalle:

1. Ejecución de una convolución transpuesta para aumentar la resolución de las características. En este paso el alto y ancho se ve duplicado.

2. Concatenación de las características del tensor f del camino de codificación (flechas grises en la Figura 3.4).

3. Empleo de una capa *dropout* del 30 %.

4. Aplicación de dos convoluciones 3x3, con *same padding* y función $ReLU$.

La salida del *decoder* presenta el mismo ancho y largo que la imagen de entrada a la red (ya sea completa o por parches), mientras que el número de canales, correspondiente a la última dimensión, se establece en 64. Por esta razón, se aplica una última convolución 1x1 con activación $ReLU$, la cual reduce el número de canales finales a 1, generando así la imagen predicha deseada. No obstante, podemos observar como en la Figura 3.4 establecemos 1 canal para los experimentos supervisados, y 2 canales para los experimentos autosupervisados. Si bien la convolución especificada en este párrafo es aplicada al supervisado porque directamente va a ser comparado con las etiquetas (post-T1w adquirida), los modelos de aprendizaje autosupervisado

a la salida de la U-Net nos reportan los mapas post-T1 y post-T2 (dos canales), y será a través de las ecuaciones teóricas de las secuencias de pulsos, un concepto físico anteriormente explicado, como obtendremos la post-T1w sintética. Concretamente aplicaremos la Ec. 2.5. Esto se verá al detalle en la Sección 3.4.

3.3.2 Arquitectura U-Net ViT

La aplicación de ViT en la arquitectura U-Net original desarrollada en el anterior apartado surge a raíz de los avances en materia de imagen que ha supuesto esta nueva tecnología. La implementación realizada se basa en el el artículo [37].

En un primer lugar, nuestro tensor de entrada, correspondiente a la salida del cuarto bloque de *downsampling* se pasa por dos convoluciones separables. Una generará nuestro tensor Q, y la otra el tensor K y V, conjuntamente, para después ser separados en dos matrices. Estos tensores son fundamentales para el mecanismo de autoatención posterior. Se explican a continuación:

- Tensor consulta Q (*query*): se utiliza para "preguntar" sobre la relevancia de las distintas posiciones de la secuencia. En nuestro caso, se comparará con las posiciones del vector K para obtener una puntuación de atención. El tensor de entrada se convierte en Q mediante la aplicación de la primera convolución separable (*DepthwiseConv2D*).

- Tensor clave K (*key*): va a representar las características de cada posición en la secuencia. Será comparado con Q, para determinar el peso que se le debe dar a cada una de las posiciones.

- Tensor valores V (*value*): es lo que realmente se va a trasladar a la salida. Contienen la información que será procesada en función de las puntuaciones de atención obtenidas de comparar Q y K. Tanto K como V se obtienen de realizar una convolución separable a la entrada original, pero de forma conjunta. Será después cuando se separen ambos tensores en dos matrices.

Tras este proceso, se realiza el cálculo de las atenciones aplicando el producto escalar entre la consulta y la clave. Este paso termina con una capa *Softmax*, que nos normaliza los coeficientes de atención. Las atenciones calculadas de este paso, serán multiplicadas a los valores, que se reordenan para obtener la salida final del mecanismo de atención. Antes de adquirir esta salida, se le aplica una convolución 1x1 que la proyecta al espacio original del que partíamos.

El resultado del proceso de atención se va a sumar a la entrada original, para formar una conexión residual. Este nuevo tensor será introducido en la siguiente y última fase del transformador.

La última parte de la tecnología ViT utilizada es un bloque *feed-forward*. Esto es, un conjunto de capas que va a procesar la información hacia adelante y de forma secuencial. La conexión residual será introducida en este bloque, con una convolución 1x1 que multiplica el número de canales de entrada por un factor (4), en primer lugar. Tras esto, se aplica una función de activación $GELU$, similar a la $ReLU$, pero con una pequeña curva para los valores menores de 0 [42], y por último, se proyecta de vuelta a la dimensión de los canales originales con una convolución 1x1. El resultado de la etapa *feed-forward* se suma a la salida original de la atención, y esta se corresponde ya con la salida de nuestro transformador.

Todos los pasos realizados se corresponden con 1 capa del transformador. Sin embargo, debemos tener en cuenta que el número de capas de transformador especificadas ha sido de 4. Por tanto, la salida del ViT será el resultado de iterar 4 veces este procedimiento, que una vez termina, pasa como entrada de la doble convolución 3x3 del cuello de botella original. Esta información, y los pasos previos explicados, desde el final del camino de contracción de la U-Net, hasta el inicio del camino de expansión, pasando por los ViT y el cuello de botella primigenio, aparecen sintetizados en la Figura 3.5.

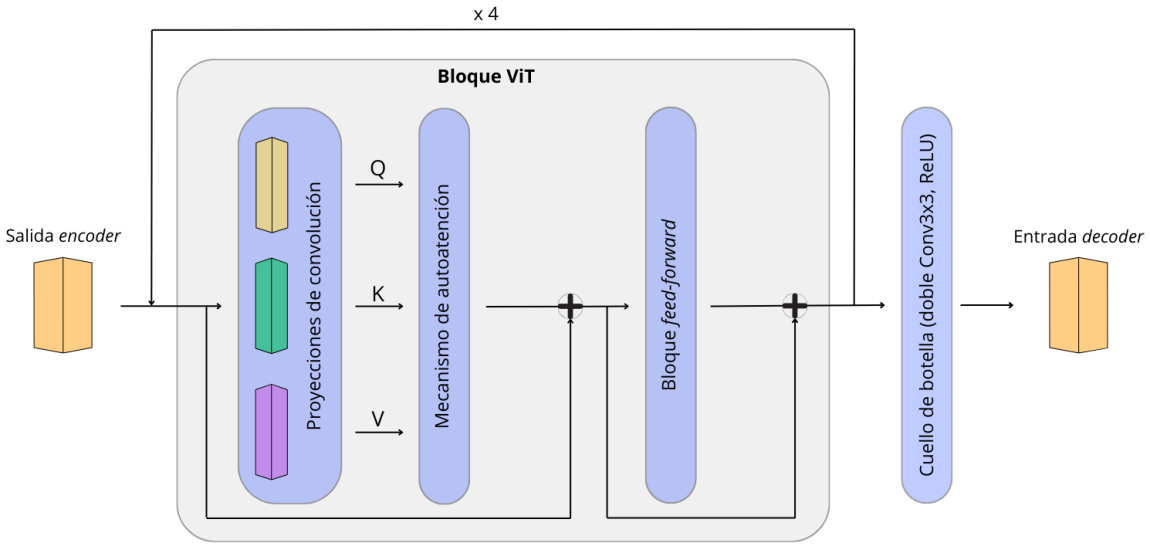


Figura 3.5: Etapas del ViT implementado en la arquitectura original.

3.4 Tipos de aprendizaje

3.4.1 Aprendizaje supervisado

Como hemos desarrollado en la sección anterior, para modelos con este aprendizaje obtenemos como salida de la red un corte bidimensional, nuestra predicción, que será directamente comparada con su etiqueta, la imagen post-T1w adquirida. Podemos visualizar el enfoque propuesto en la primera parte de la Figura 3.6.

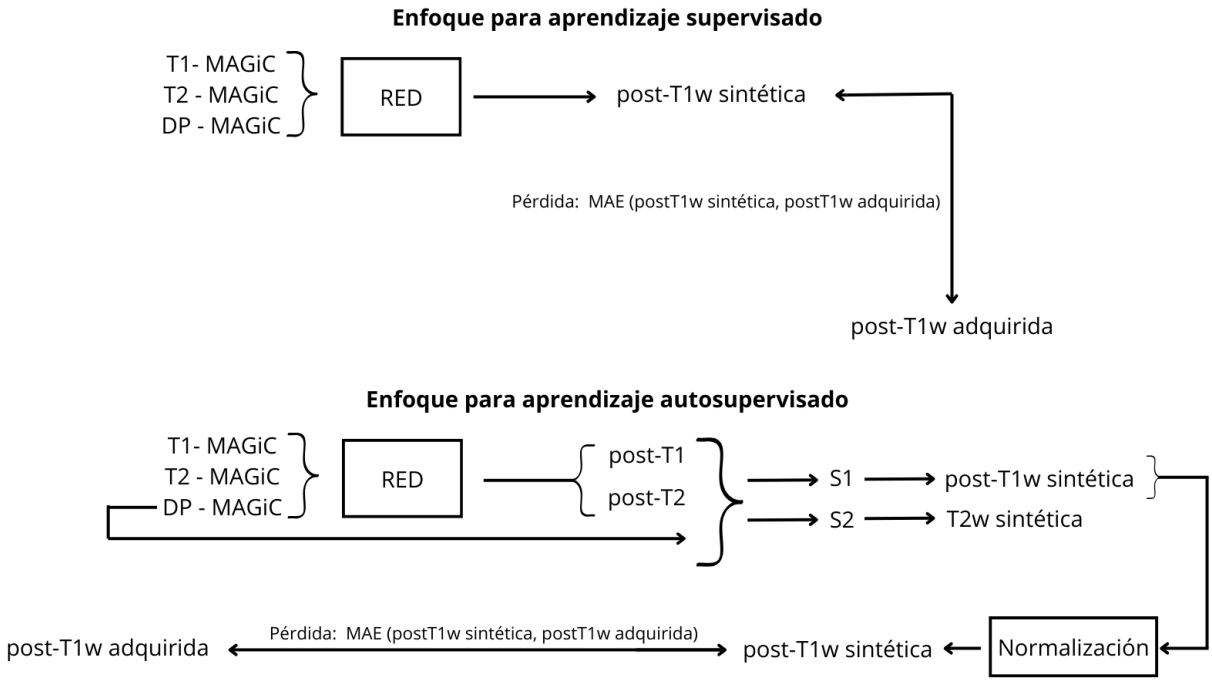


Figura 3.6: Esquema del planteamiento seguido para ambos aprendizajes.

La métrica de pérdidas utilizada para la optimización del modelo es el error absoluto medio (MAE, de sus siglas en inglés *mean absolute error*), calculado entre la imagen post-T1w obtenida con la red y la correspondiente imagen adquirida. Se ha utilizado la misma métrica en ambos aprendizajes (véase Figura 3.6).

3.4.2 Aprendizaje autosupervisado

En relación con el aprendizaje autosupervisado, la salida de nuestra red, tanto U-Net como U-Net ViT, son los mapas post-T1 y post-T2. Estos mapas, junto con el mapa DP original, son aplicados en la ecuación de la secuencia de pulsos 2.5, para la obtención de las imagen post-T1w sintética.

El esquema presente en la Figura 3.6 muestra el experimento que más se ha trabajado durante el proyecto para este aprendizaje. En él, tras la aplicación de la ecuación teórica 2.5, se aplica una normalización. Como la imagen post-T1w adquirida requería de una normalización, y a los mapas aplicados solo se les realiza un cambio de escala, debemos implementar la misma normalización a la salida de nuestra capa *lambda* de aplicación de la ecuación 2.5 (véase Figura 3.6). Es tras este paso, cuando se produce la introducción de la etiqueta y la optimización del modelo a partir de la métrica de pérdidas.

El objetivo de los experimentos de aprendizaje autosupervisado ha sido siempre la predicción de una correcta imagen post-T1w. Sabemos que incorporar conceptos físicos ayuda a cumplir este objetivo. No obstante, parte de este proceso implica la correcta visualización de los mapas post-T1 y post-T2, obtenidos previos a las ecuaciones. Una visualización correcta de estos mapas sería interesante y aportaría valor añadido al procedimiento.

El enfoque planteado en la Figura 3.6 podría no otorgar unos mapas con valores realistas, por lo que para intentar solventar este problema, se realizó un experimento a mayores. Este experimento consistió en la adición de la imagen T2w al modelo general. De esta forma, tras la aplicación de la Ec. 2.6, pudimos obtener la T2w sintética, que tras un proceso de normalización paralelo al realizado con la post-T1w, nos permitió compararla con la imagen T2w adquirida para el cálculo de la función de pérdidas. Con esto, conseguimos que la optimización del modelo se realice con dos condicionantes simultáneos, es decir, dos conjuntos de etiquetas y dos métricas de pérdidas.

3.5 Reconstrucción de las parches

La imagen completa se predice de forma sencilla para los experimentos de imagen completa, y después se almacena en formato *NIfTI* comprimido (*.nii.gz*). En relación a los experimentos basados en parches, debemos introducir una función de reconstrucción de los parches predichos. Esta función se puede definir en dos fases:

- Una primera etapa donde se realiza la concatenación total de los parches, sumados uno detrás del otro, para cada uno de los *stacks*, y en su debido orden. Esto se realiza avanzando en el eje horizontal y vertical con un paso del tamaño del parche menos el tamaño del solapamiento (64-16). La construcción final de la Figura 3.3, aunque no pensada para este formato, puede sernos útil para entender la configuración de esta fase.

- Una segunda etapa de normalización de la reconstrucción. Como se realiza la suma, las regiones solapadas tendrán unos valores que deben ser normalizados, bien por 2, si la región solapada pertenece a dos parches, o por 4 si es el resultado en esa región es la suma de 4 parches. Podemos visualizar el enfoque de este normalizador en la Figura 3.7.

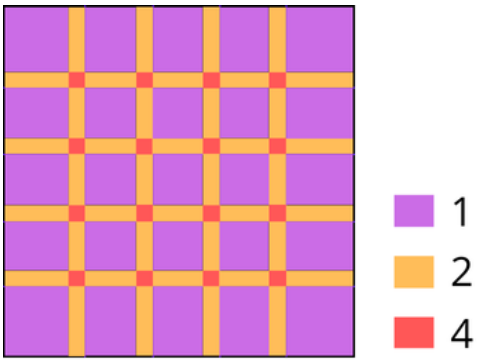


Figura 3.7: Concepto de normalizador aplicado, con los distintos valores divisores.

3.6 Experimentos

Con una visión general de las distintas formas de entrenamiento, red y entrada, definimos las pruebas realizadas durante el trabajo:

1. **U-Net s+ic** : arquitectura **U-Net** con entrenamiento **supervisado** de **imagen completa**.

2. **U-Net as+ic**: arquitectura **U-Net** con entrenamiento **autosupervisado** de **imagen completa**.
3. **U-Net as+ic+d**: arquitectura **U-Net** con entrenamiento **autosupervisado** de **imagen completa** y **doble salida** (post-T1w y T2w).
4. **U-Net s+p**: arquitectura **U-Net** con entrenamiento **supervisado** basado en **parches**.
5. **U-Net as+p**: arquitectura **U-Net** con entrenamiento **autosupervisado** basado en **parches**.
6. **U-Net as+p+aug**: arquitectura **U-Net** con entrenamiento **autosupervisado** basado en **parches** con **aumento de datos** de parches de regiones con realce o edema tumoral.
7. **U-Net ViT s+p**: arquitectura **U-Net ViT** con entrenamiento **supervisado** basado en **parches**.
8. **U-Net ViT as+p**: arquitecutra **U-Net ViT** con entrenamiento **autosupervisado** basado en **parches**.

En este párrafo se expone la justificación de la no realización de todos los experimentos posibles, para todas las combinaciones. En primer lugar, la imagen de doble salida solo se puede aplicar a modelos de aprendizaje autosupervisado, puesto que son los que generan los mapas post-T1 y post-T2. A mayores, solo se aplicó a la imagen completa porque la tecnología por parches, incluso con solapamiento, generaba efecto de bloques en los mapas post-contraste resultantes. Por otro lado, como se ha expuesto en la Sección 3.2, la realización del experimento de aumento de datos con parches de regiones con realce se planteó como potencial mejora del mejor experimento que se tenía hasta ese momento (autosupervisado por parches). No se ha extrapolado a otro tipo de experimentos porque acarrea mucha carga computacional, y porque como veremos en los resultados, no produjo mejoras significativas en esta primera prueba. Por último, se debe destacar que no se han realizado experimentos U-Net ViT de imagen completa por problemas de carga computacional (más de 45 millones de parámetros y tensores muy grandes en algunas capas), y por problemas de convergencia y optimización asociados. En adición, la tecnología de transformadores suaviza mucho el efecto de bloques respecto a los modelos de la arquitectura original (incluso sin aplicar el solapamiento), enfocando así su aplicación a la entrada por parches.

Los ensayos definidos, han sido realizados para cada paciente, siguiendo la técnica de validación cruzada LOO. El número de pacientes de entrenamiento se ha fijado en 11, mientras que el de validación en 2 (el restante es nuestro test). La división entre entrenamiento y validación fue aleatoria, garantizando siempre que uno de los pacientes de validación presentara regiones con realce, y el otro no.

3.7 Evaluación

La metodología de evaluación de resultados seguida podemos dividirla en una comprobación visual, y en una comprobación mediante métricas de calidad.

Las métricas utilizadas han sido 3. Todas ellas tienen por motivo, evaluar la calidad de las imágenes, comparando la predicción del volumen 3D de cada prueba realizada, con su etiqueta. Las definimos:

- **Índice de similitud estructural, SSIM** (de sus siglas en inglés *structural similarity index measure*): es una métrica compleja, que no opera como las métricas comunes midiendo las diferencias absolutas de los píxeles. La SSIM va a estar más relacionada con la forma (y complejidad) con la que los humanos perciben la calidad visual, a partir de la medición, por separado, de la luminancia (brillo de la imagen), el contraste (variaciones en la intensidad de los píxeles), y la estructura (cambios de intensidad en patrones locales) [43]. Oscila entre 0 y 1, siendo 1 el mejor valor posible.
- **Relación señal a ruido máxima, PSNR** (de sus siglas en inglés *peak signal-to-noise ratio*): define la relación entre la potencia de máxima de una señal, y la potencia del ruido que afecta a la imagen predicha, en comparación con la adquirida [44]. Cuanto mayor sea esta métrica, mejor imagen predicha.
- **Error cuadrático medio, MSE** (de sus siglas en inglés *mean squared error*): mide el promedio de los errores al cuadrado de nuestras predicciones, píxel a píxel. Al contrario que la SSIM y la PSNR, un mayor valor del MSE implicará una peor predicción por parte de nuestro modelo (una mayor diferencia media entre los píxeles de la imagen original y la predicha).

Estas tres métricas han sido calculadas para dos criterios, enfocados en el análisis del volumen predicho y su etiqueta, resultando en un total de 6 métricas obtenidas. A continuación, los describimos:

- **Volumen completo sin fondo**. Para ello, previo al cálculo de las métricas de calidad, se procedió a la retirada de los valores nulos de la imagen etiquetada, siendo extrapolado a la predicha.
- **Volumen de la región de la máscara, sin fondo**. Esta máscara presenta valor 1 para regiones con edema, valores 2 para zonas con realce, y valores 0 para el resto. Se ha procedido a la unión de todo el tejido tumoral, convirtiendo los valores 2 a 1. Al volumen predicho y etiquetado se les multiplica por esta máscara, y después se realiza la anulación en el cálculo de los píxeles nulos (fondo) de la imagen procesada.

Capítulo 4

Resultados

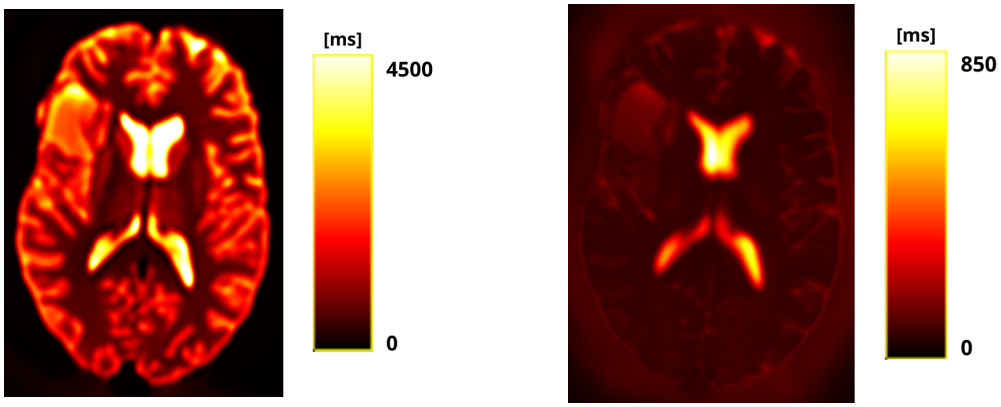
En esta sección del trabajo se presentan y discuten los resultados obtenidos en los 8 experimentos, mediante el método aplicado en el capítulo 3. Se realizan dos tipos de análisis de resultados, que se exponen a continuación.

4.1 Por exposición visual

En primer lugar, se realiza una inspección visual de los mapas post-T1 y post-T2, obtenidos gracias al experimento de doble salida enfocado a ello. Posteriormente, se muestran las máscaras de los pacientes seleccionados para la exposición de los resultados, y en última instancia, se disponen las imágenes post-T1w sintéticas de todos ellos, para todos los experimentos, junto con la post-T1w adquirida o *groundtruth*.

En la Figura 4.1 observamos los resultados de los mapas post-contraste para el experimento de doble salida. Se recoge como ejemplo los resultados para el paciente 2, que también es sujeto de exposición más adelante con la visualización de la post-T1w sintética (no el mismo corte). El experimento U-Net as+ic+d permite la obtención de unos mapas con valores realistas, que no se han podido lograr con ningún otro experimento.

En la Figura 4.2, podemos visualizar la segmentación de la máscara para el corte en cuestión de las predicciones sintéticas posteriores. El examen de estas máscaras nos permiten una mejora en la comprensión del realce obtenido entre experimentos, actuando de soporte de la comparativa realizada.

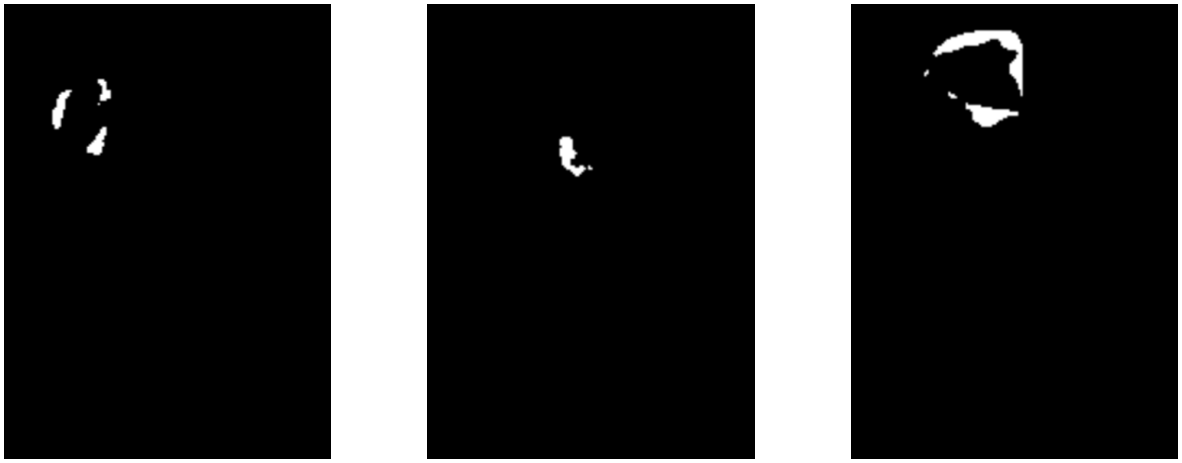


Mapa post-T1 Mapa post-T2

Figura 4.1: Mapas post-T1 y post-T2 predichos con el experimento U-Net as+ic+d, del paciente 2.

Para finalizar el apartado visual, contemplamos la predicción de la post-T1w sintética para el paciente 2, 4, y 14 de nuestra base de datos. Lo examinamos a partir de las Figuras 4.3, 4.4 y 4.5, respectivamente. En ellas, podemos encontrar, en primer lugar, la imagen post-T1w adquirida, y acto seguido, el resultado de los 8 experimentos en el orden planteado en la Sección 3.6. Naturalmente, para cada paciente las 9 imágenes y la máscara correspondiente se han adquirido en el mismo corte. En adición, todas las imágenes de un mismo paciente han sido obtenidas en un mismo intervalo de valores, fijando para cada imagen el mínimo y máximo de la imagen adquirida.

Entrando en el análisis, observamos como para el paciente 2 (Figura 4.3), el realce tumoral, o al menos una parte de él, se predice aceptablemente por la mayoría de experimentos. Concretamente, el ubicado en el margen izquierdo de la imagen. Esta afirmación (y todas), han sido realizadas solo después de examinar la máscara de la Figura 4.2. A simple vista, el experimento U-Net as+ic es el que más se acerca a la imagen adquirida, y aparentemente, a la predicción del realce. Los experimentos basados en imagen completa logran unas predicciones generales de la imagen notorias. No obstante, sufren en la predicción cuando se trata de áreas concretas. En este caso, a pesar de que la imagen U-Net as+ic es buena en líneas generales, incluida la región del tumor y sus alrededores, el realce es más significativo en los modelos de U-Net ViT. Concretamente, es el U-Net ViT as+p el que alcanza unas tasas de realce (intensidad) más cercanas a la post-T1w adquirida, y más precisas respecto la máscara. Por otro lado, la zona de resección tumoral es predicha por el U-Net ViT s+p mejor que por cualquier otro experimento.



Máscara paciente 2 Máscara paciente 4 Máscara paciente 14

Figura 4.2: Máscaras del realce y edema tumoral de los pacientes utilizados en la visualización.

Con el paciente 4, observamos una región de realce más pequeña y compleja, ubicada en una zona más interna del parénquima cerebral. En la Figura 4.4 constatamos como la predicción, en esta ocasión, es de peor calidad. A pesar de ello, es interesante visualizar la mejora de realce a medida que aumenta la complejidad de los experimentos. Con este sujeto, solo el modelo U-Net ViT as + p (el último) consigue una predicción de realce clara y brillante, muy próxima a la de la máscara expuesta, pudiendo ser esto indicativo del gran valor de este último experimento.

En último lugar, se ubica el paciente 14 (Figura 4.5). Para este sujeto, es el experimento U-Net as+p el que predice, con mucha holgura sobre el resto, la región de realce tumoral. Este paciente es un caso extraño porque, si nos fijamos, la máscara abarca áreas pegadas a la región extirpada, que en las predicciones son valores negros. Esta característica se ve, sobre todo, en la imagen adquirida, donde existen distintas tonalidades dentro de la zona de resección, que no aparecen en ninguna predicción. Esta es la principal causa de las malas métricas de MSE que tenemos para este paciente, y que se exponen en la Sección 4.2. A pesar de ello, los resultados del experimento autosupervisado por parches son alentadores, y han sido el motivo de incorporación del experimento con aumento de datos en el trabajo, a pesar de que no ha otorgado mejoras, finalmente.

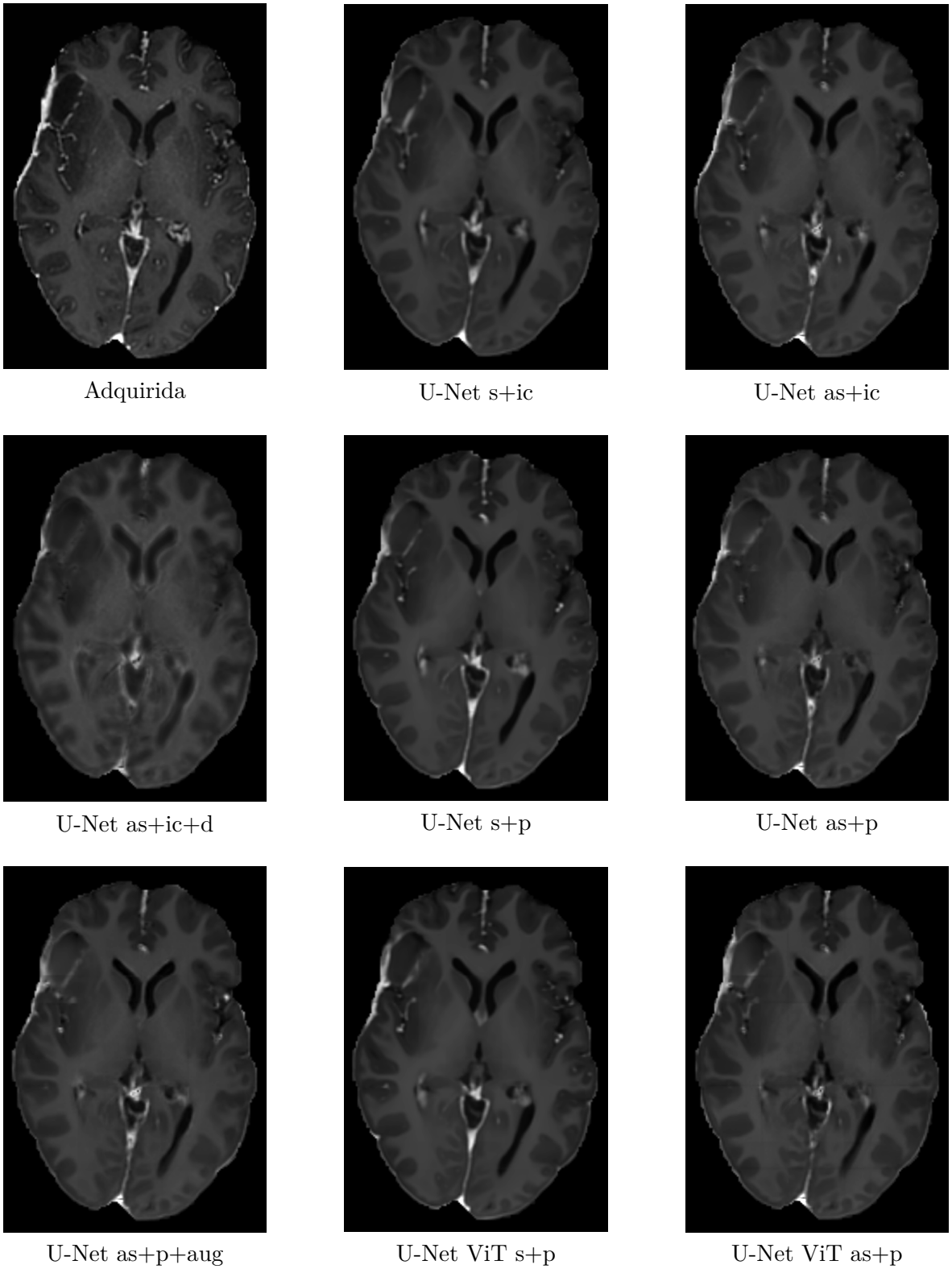


Figura 4.3: Predicción de post-T1w sintética para el paciente 2.

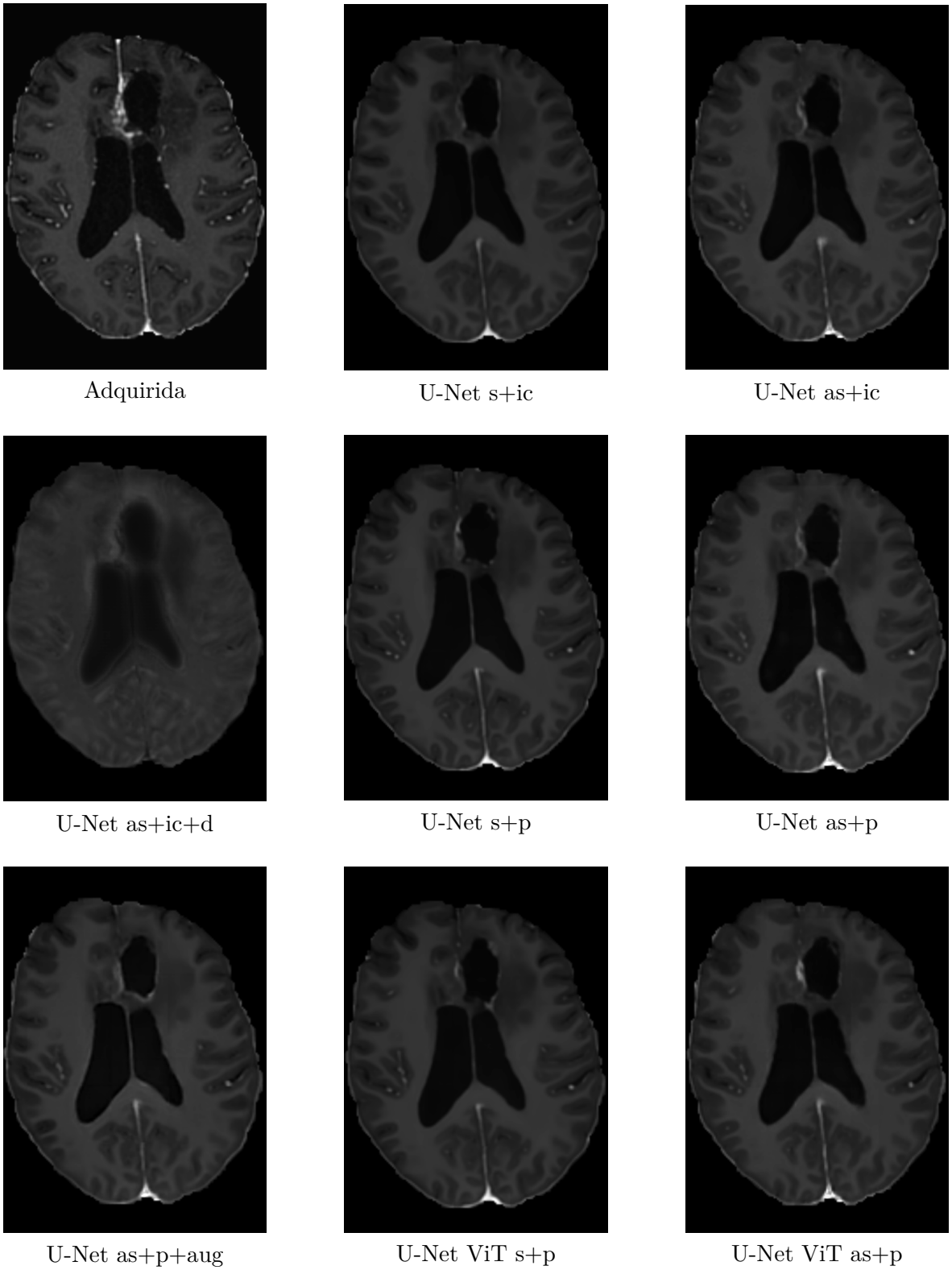


Figura 4.4: Predicción de post-T1w sintética para el paciente 4.

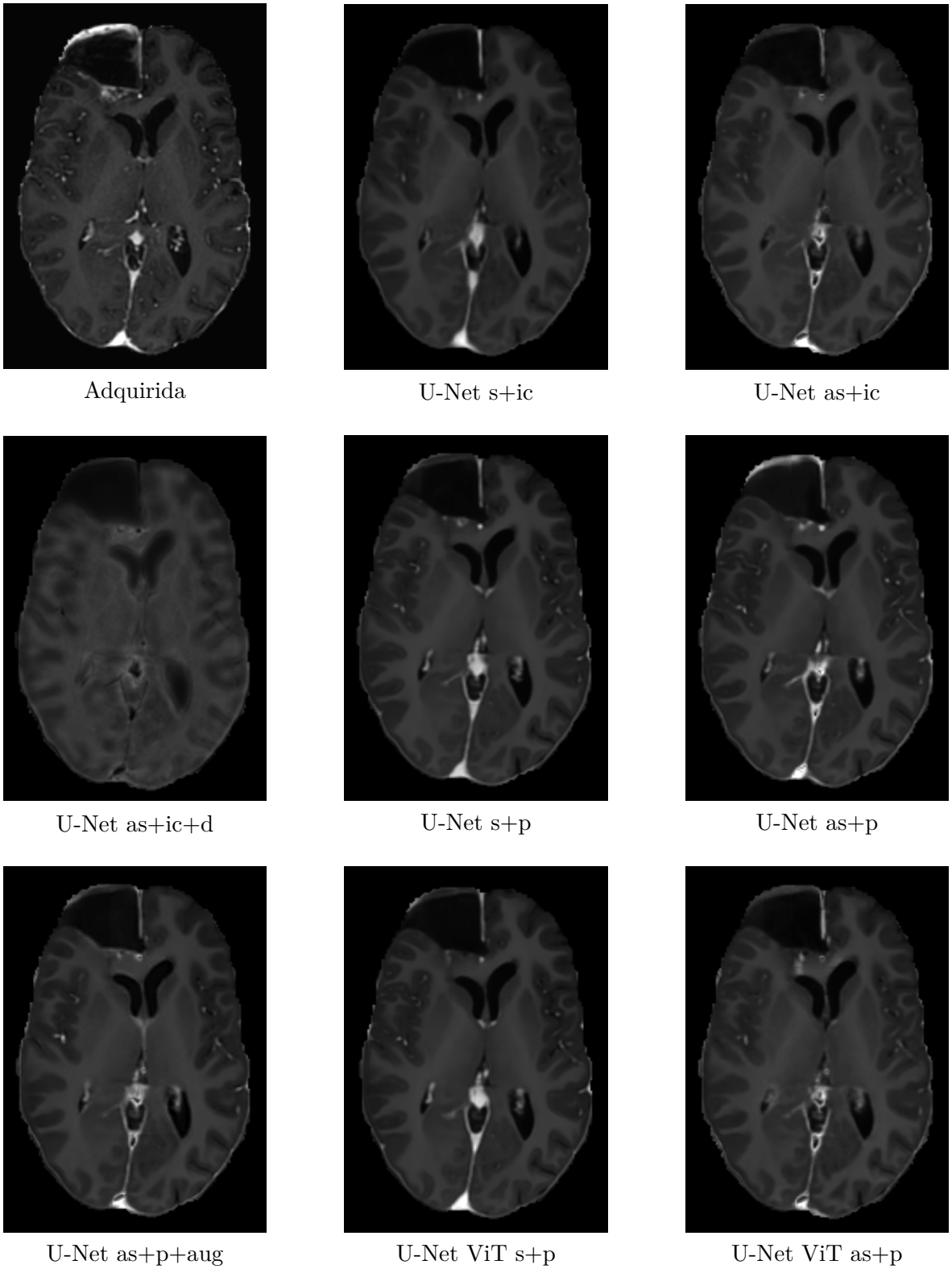


Figura 4.5: Predicción de post-T1w sintética para el paciente 14.

4.2 Por métricas de calidad

Se introducen los resultados de las métricas de calidad planteadas en la Sección 3.7 en forma de diagramas de caja o bigotes (comúnmente conocidos como *boxplots*), y se procede a su análisis. En este caso, cada uno de los diagramas representado se genera a partir de los 14 valores de la métrica obtenida de cada paciente (LOO), para el experimento en cuestión. Los diagramas de caja son un tipo de representación de la distribución de un conjunto de datos determinado, y podemos descomponerlos en:

- Una línea horizontal interna en la caja. Representa la mediana, es decir, la mitad de los datos se encontrarán por debajo de esta franja, y la otra mitad por encima.
- Una caja que define el 50 % de los datos centrales (rango intercuartílico, IQR). El IQR viene dado por el borde inferior de la caja, que representa el primer cuartil ($Q1$), que indica que el 25 % de los valores se encuentran por debajo, y el tercer cuartil ($Q3$), correspondiente al borde superior delimitante de la caja, que nos indica que el 75 % de los datos totales se encuentran por debajo de este borde.
- Los bigotes del diagrama. Extendidos desde los límites de la caja, buscan aproximarse a los datos mínimos y máximos, pero solo hasta un cierto rango inferior y superior ($Q_1 - 1{,}5 \times \text{IQR}$ y $Q_3 + 1{,}5 \times \text{IQR}$, respectivamente). Valores más allá de los bigotes, se consideran atípicos.
- Valores atípicos (*outliers*): son los datos más alejados que el límite inferior y superior de los bigotes.

En la Figura 4.6 se muestran las métricas *SSIM*, *PSNR*, y *MSE* para la imagen completa sin fondo, obtenidas para los 8 experimentos. El orden de aparición de los mismos en cada una atiende al orden numérico establecido en el apartado de evaluación del Capítulo 3. En un primer lugar, establecemos los experimentos para la arquitectura U-Net de imagen completa (ic): supervisado (U-Net s+ic), autosupervisado (U-Net as+ic) y autosupervisado de doble salida (as+ic+d). Lo siguiente es la U-Net con entrada basada en parches (p) supervisado (U-Net s+p), autosupervisado (U-Net as+p), y autosupervisado con aumento de datos (U-Net as+p+aug). Los dos últimos se corresponden con la arquitectura que incorpora un ViT a la U-Net original utilizada (U-Net ViT). Siguiendo el orden de aprendizajes, primero se encuentra el supervisado por parches (U-Net ViT s+p), y en último lugar el autosupervisado por parches (U-Net ViT as+p). No se han realizado experimentos U-Net ViT de imagen completa, debido a su alto coste

computacional (superior a 45 millones de parámetros y tensores muy grandes en algunas capas), y a problemas de convergencia y optimización asociados.

Con la SSIM y PSNR de la Figura 4.6, se observa una mejora de rendimiento considerable con la utilización del aprendizaje autosupervisado (as), respecto al supervisado, por lo que deducimos que la aplicación de conceptos físicos como la ecuación teórica de la secuencia de pulsos ayuda a la optimización de nuestros modelos. Si bien la mejora con este aprendizaje se conserva desde el experimento más simple (as+ic), las pruebas con aprendizaje supervisado nos permiten corroborar que la tendencia de las métricas de calidad es ascendente con la aplicación de parches, y de la arquitectura ViT aplicada a parches. Por otro lado, tenemos el experimento con la doble salida (d), que no resulta en unas buenas predicciones de la post-T1w (y por tanto, en estas métricas de calidad), pero sí mejora sustancialmente los mapas post-T1 y post-T2, como podemos ver en la exposición visual. El resto de experimentos de aprendizaje autosupervisado, a pesar de ser, con diferencia, nuestros mejores experimentos, no predicen unos mapas con un sentido físico realista. Los ViTs merecen una mención especial también, puesto que han permitido una mayor consistencia en la variación de la PSNR entre pacientes (nótese el grosor de la caja U-Net Vit as+p, con el resto de as) y, en general, una mejor predicción del realce de las imágenes.

La última métrica expuesta en la Figura 4.6 es el MSE del volumen completo. En él, contra todo pronóstico, son los modelos supervisados los que obtienen mejores resultados. Esto hace pensar que, aunque la evaluación píxel a píxel sea más positiva para estos experimentos, cuando se introducen métricas más profundas y robustas como la SSIM, los resultados cambian. Por esta capacidad limitada, y sumado a que la diferencia es muy pequeña contando la intensidad total que presentan nuestras imágenes, se concluye que la MSE presenta una peor evaluación de rendimiento que los diagramas de las otras dos métricas. A pesar de ello, debemos destacar la reducción del error que influyen los parches a nuestros experimentos supervisados.

En lo referido a la Figura 4.7, se sigue el mismo procedimiento que para la Figura 4.6, pero con el cálculo de las métricas para la región de la máscara previa retirada de los valores de fondo. Los resultados de estas métricas son más pobres que para la imagen completa, debido a la escala de las mismas en el margen izquierdo. Además, el tamaño elevado de las cajas en los diagramas nos indica que el modelo es menos preciso entre experimentos, por la distinta complejidad que

presenta la predicción en las zonas con realce, entre unos pacientes y otros. En consecuencia, los datos obtenidos para cada métrica se encuentran más alejados entre ellos de lo que podemos corroborar con la Figura 4.6 para imagen completa.

Con las métricas de SSIM y PSNR no se pueden alcanzar conclusiones significativas, si bien es cierto que la prueba de aumento de datos (aug), realiza una predicción más cercana entre pacientes, al presentar la caja más pequeña de entre todos los diagramas. El modelo de doble salida, como podíamos imaginar, es también el peor en la predicción de las regiones con realce y edema. El que posee una mejor mediana sigue siendo el ensayo U-Net ViT as+p, sin embargo, tan solo nos da un valor próximo a 0'55 para la SSIM.

Con el MSE, existe una particularidad importante. Se puede observar un valor atípico para cada una de las 8 pruebas realizadas. Este *outlier*, se corresponde en todas ellas con el MSE de la región de la máscara del paciente 14. El paciente 14 tiene localizado el tumor en el lado izquierdo del lóbulo frontal, próximo al fondo negro de la imagen, como podemos ver en la Figura 4.5. Además, se ha procedido quirúrgicamente a su extirpación, por lo que podemos verlo totalmente negro (no hay parénquima cerebral). El realce de este paciente por encima del tumor se desarrolla en el borde de la misma imagen con el fondo, quedando, por tanto, ubicado entre el tumor (valores nulos por extirpación), y el fondo (valores nulos). Esto nos hace pensar que una mala segmentación de la máscara, o una mala predicción de nuestra imagen, reporte diferencias muy grandes de valores, debido a que donde se espera un píxel con realce (valores de intensidad de 6-7), quizá se haya predicho o se encuentre un píxel negro (valor 0), afectando contundentemente al MSE. Observando minuciosamente la máscara obtenida por *HD-GLIO* y la imagen adquirida para este paciente (véase Figura 4.2), establece píxeles con realce o edema en lugares donde la predicción ha indicado que es región extirpada. Siguiendo este hilo, podemos observar como el experimento U-Net as+p otorga el MSE con el valor atípico menos alejado de todos (próximo a 3, sin superarlo). Este experimento coincide con la mejor predicción realizada para el paciente 14 (véase Figura 4.5), por lo que, en efecto y como no podía ser de otra manera, existe una correlación entre ambos hechos.

Finalizando con las conclusiones del MSE de la Figura 4.7, se observa como el 50 % de los datos se encuentra compactado en valores bajos del MSE (la mediana está muy por debajo de la

mitad de la caja), mientras que los otros valores por encima del nivel medio, se encuentran mucho más dispersos. Esto es así porque el MSE posee una mayor variación en la predicción cuando los pacientes presentan zonas de realce complejas, que además son muy distintas entre ellas. Con pacientes cuyas regiones de realce y edema tumoral son más cómodas de predecir, los valores de MSE obtenidos están más próximos entre ellos, y por ello se ubican en una región tan pequeña del diagrama (mitad inferior desde la mediana). Si bien observábamos una tendencia de los experimentos supervisados en el MSE de volumen completo, en lo referido a las regiones de realce, las pruebas autosupervisadas consiguen mejores resultados para esta métrica, en especial el autosupervisado por parches. Esto está directamente relacionado con la exposición visual realizada.

Por otro lado, los diagramas de caja al trabajar con los datos exactos, y con su dispersión, son más robustos ante valores atípicos. En el lado opuesto se encuentra la media aritmética, que a pesar de ser más sensible ante estos *outliers*, se ha considerado también importante para este trabajo y se ha introducido una tabla para su comparación entre experimentos. Esta Tabla 4.1 nos indica el valor promedio de cada métrica (columna), para cada experimento (fila). El último y más complejo experimento obtiene los mejores resultados, por lo que se confirma que el uso de modelos físicos y arquitecturas de red avanzadas como los ViT, mejoran la calidad de las predicciones. Por otro lado, la arquitectura U-Net ViT supervisada por parches otorga los mejores resultados de MSE, en lo referido a la media de los 14 valores para el volumen completo sin fondo, y el aprendizaje autosupervisado por parches para la arquitectura U-Net original, en la región de realce y edema tumoral.

	SSIM	**PSNR**	**MSE**	**SSIM masc**	**PSNR masc**	**MSE masc**
U-Net s+ic	0.79057	25.59151	0.18095	0.54165	18.17563	0.83359
U-Net as+ic	0.83096	28.19276	0.19015	0.53387	17.90367	0.70688
U-Net as+ic+d	0.80048	26.18810	0.15649	0.55000	18.19416	0.74147
U-Net s+p	0.74051	24.01492	0.24917	0.47887	17.46725	1.07084
U-Net as+p	0.83072	28.21894	0.18458	0.54855	18.26496	**0.65124**
U-Net as+p+aug	0.80789	26.96604	0.19076	0.54149	17.87024	0.74549
U-Net ViT s+p	0.80344	26.38644	**0.15204**	0.55046	18.27350	0.72597
U-Net ViT as+p	**0.83302**	**28.28715**	0.19733	**0.55487**	**18.35470**	0.71666

Tabla 4.1: Tabla de la media aritmética de las métricas para cada experimento. Nótese que los valores mejores de cada métrica, han sido resaltados dentro de la tabla.

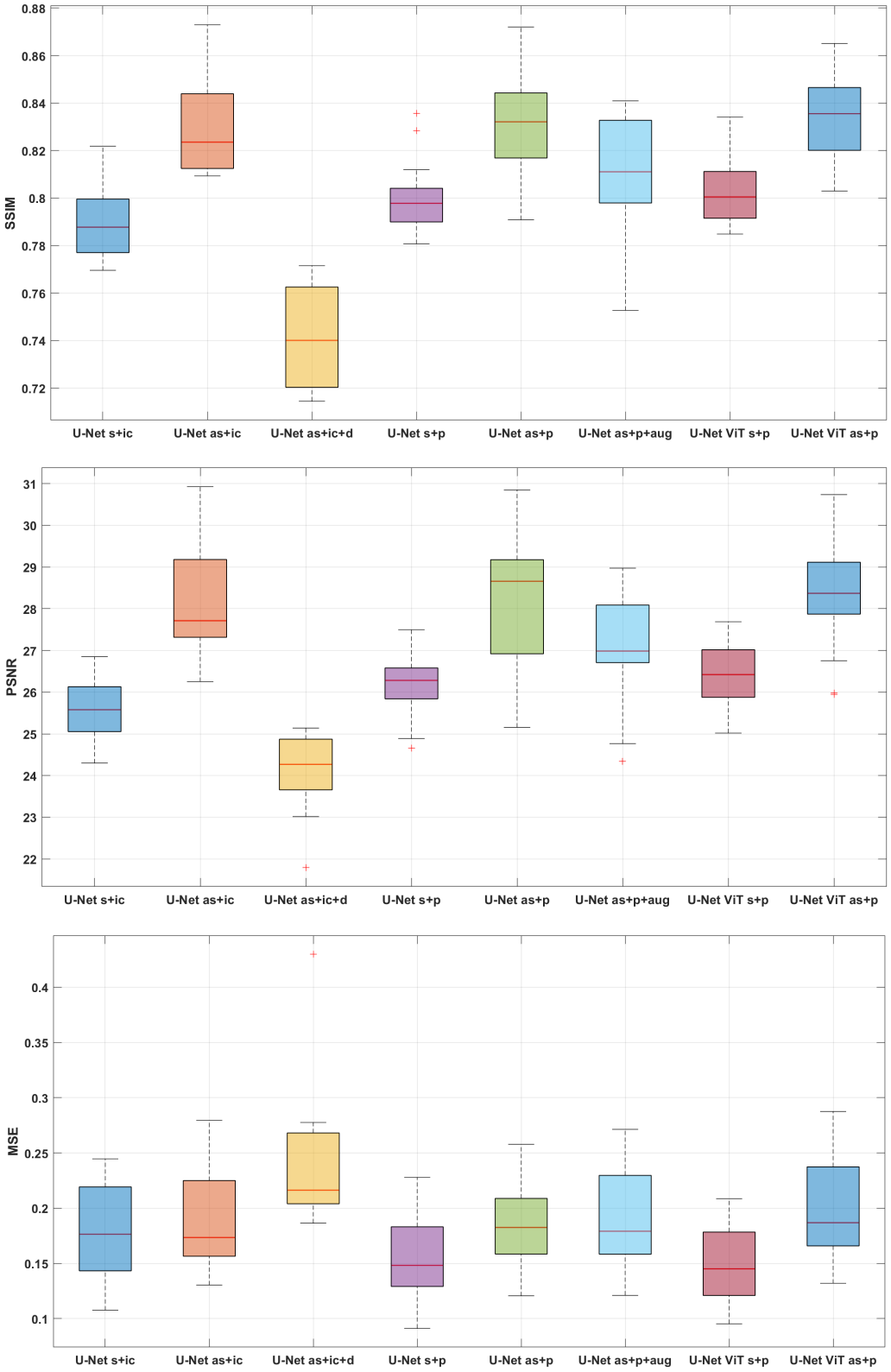


Figura 4.6: Métricas SSIM, PSNR y MSE para el volumen completo sin fondo.

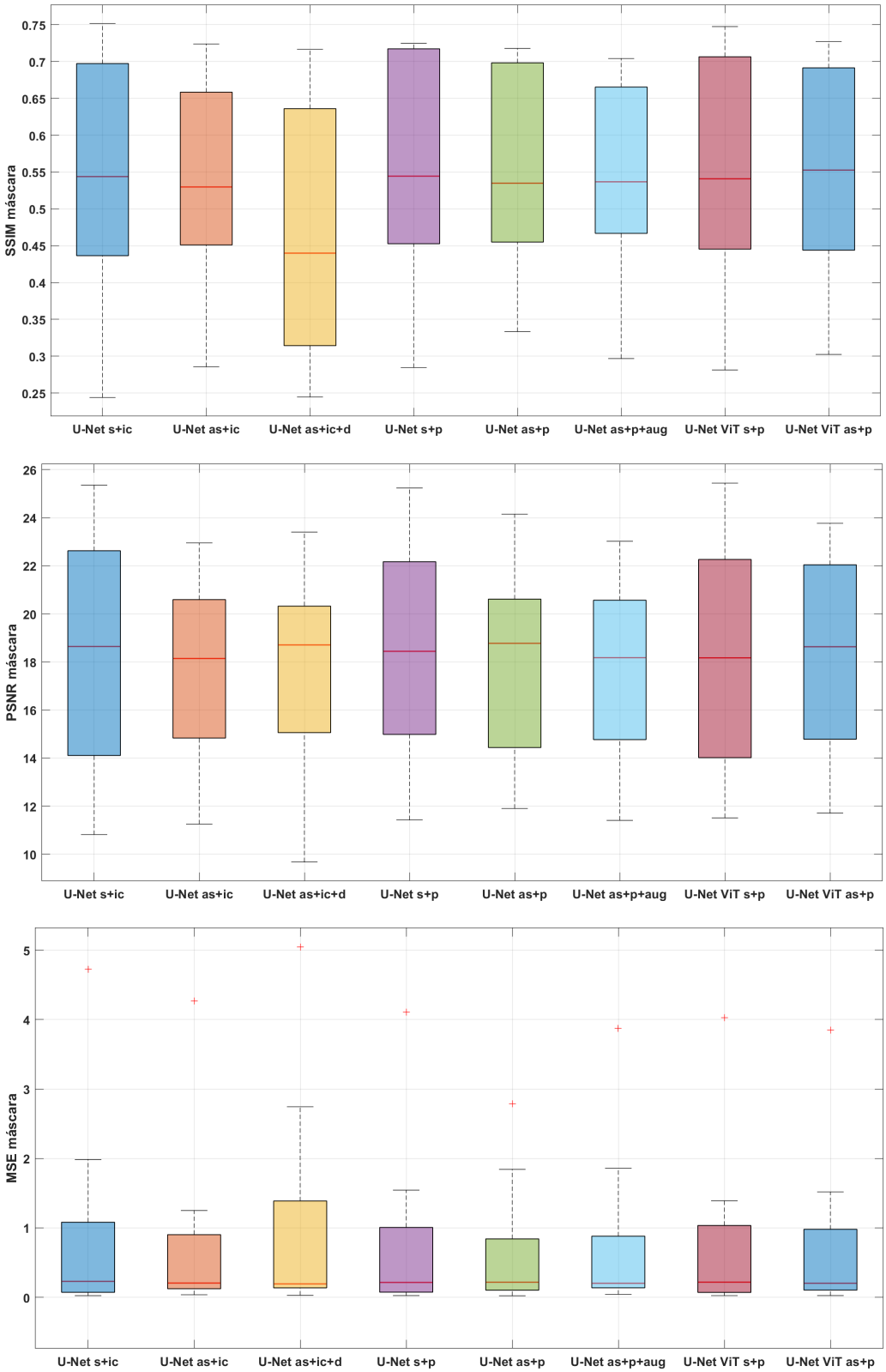


Figura 4.7: Métricas SSIM, PSNR y MSE para el volumen de la región de la máscara, sin fondo.

Capítulo 5

Conclusiones

5.1 Conclusiones

En este Trabajo de Fin de Grado, se han abordado la exploración y aplicación de nuevas técnicas de aprendizaje profundo, que nos permitieran sintetizar imágenes post-contraste a partir de adquisiciones que no hacen uso de agentes de contraste. Para ello, se han puesto en marcha distintos modelos a los que se les han ido aplicando nuevas incorporaciones, en forma de conceptos físicos, entradas enfocadas en percibir el detalle, aumento de datos, y arquitecturas de red avanzadas, como los ViT.

El objetivo principal que se ha perseguido con este proyecto ha sido la mejora en el diagnóstico de tumores cerebrales, a partir de la obtención de una imagen post-T1w sintética de calidad que nos permitiera visualizar regiones relevantes para el diagnóstico, como las áreas de realce tumoral. En el trabajo realizado, a pesar de haber logrado ostensibles avances y de introducirse nuevas integraciones que sustentan el hilo conductor del proyecto, estamos abordando un problema complejo, y las predicciones presentan todavía un margen de mejora no despreciable. En especial, cuando tratamos con áreas de realce ubicadas en zonas complejas del cerebro (pegado a otras estructuras e internamente, entre medias de tejidos de distinta composición, ...).

La principal explicación a la materia expuesta en el anterior párrafo podría ser el tamaño de nuestra base de datos. Partimos de un conjunto de 14 pacientes, y un aumento en el número de ellos podría venir acompañado de un mayor rendimiento y optimización de nuestros modelos. Otro factor limitante sería la arquitectura U-Net original. Nos basamos en una arquitectura muy sencilla, y aunque los aportes introducidos en los experimentos resulten positivos

y prometedores, podrían ser aún mejores si se trabaja con redes más avanzadas y complejas. El ejemplo más claro es el experimento por aumento de datos, donde hemos comprobado que no se da ninguna mejora respecto el experimento original sin ellos (U-Net as+p). Esto hace pensar que la arquitectura utilizada alcanza su máximo rendimiento previo a la introducción de nuevas incorporaciones, en muchas ocasiones.

Por otro lado, con el método de solapamiento de creación de parches, y posterior reconstrucción de los mismos, se sigue observando un ligero efecto "de bloques", que si bien es cierto nada tiene que ver con el original de parches exactos, sigue estando presente. Este efecto conseguimos opacarlo de forma considerable, a costa de un aumento de un efecto "borroso" en las regiones solapadas, que aparecía en menor medida con la tecnología de entrada de datos inicial planteada. De cualquier modo, los resultados visuales son mucho mejores que cuando empezamos, pero es una limitación a tener en cuenta.

También podemos destacar el caso del experimento de doble salida, y en general todo lo envolvente a la predicción de los mapas post-T1 y post-T2, obtenidos previo a las ecuaciones 2.5 y 2.6. Nuestros experimentos no se han enfocado en la predicción de estos mapas, pero sabemos que una visualización correcta de los mismos sería interesante y aportaría valor añadido al procedimiento. El planteamiento del experimento de doble salida es de gran valor, otorgándonos unos mapas post-contraste razonablemente buenos, que sin embargo, no son acompañados por una predicción realista de la post-T1w sintética. Pensamos que la falta de optimización de la post-T1w sintética puede deberse a la introducción de dos condicionantes, lo que hace que nuestros modelos ya no puedan enfocarse en la optimización de la única vía posible, sino que debe lidiar con dos convergencias simultáneas (post-T1w y T2w), generándose estos problemas.

En resumen, este Trabajo de Fin de Grado ha demostrado que el uso de modelos físicos y arquitecturas avanzadas como los ViT mejoran la calidad de las predicciones de imágenes post-T1w, en particular, de regiones con realce, sin necesidad de GBCAs. Además, el método aplicado no aumenta el tiempo de adquisición tradicional, gracias al uso de técnicas de DL. Todo ello, en aras de ayudar en el diagnóstico de gliomas.

5.2 Líneas futuras

Por todo lo expuesto en la Sección 5.1, establecemos en este apartado posibles líneas de trabajo futuro para resolver las limitaciones planteadas y posiblemente mejorar la predicciones obtenidas.

En relación a la arquitectura utilizada en el trabajo, sería muy valorable aplicar, con estas mismas incorporaciones probadas (parches, bloque ViT, ...), redes neuronales con arquitecturas más complejas, que puedan tener un mayor margen de mejora ante distintas situaciones, como es el caso que trabajamos. Sería ideal disponer de conocimientos que permitieran alcanzar una mayor comprensión de los enfoques que hay actualmente para la síntesis de imagen, y de la simbiosis, a través de estudios, que puedan tener distintas arquitecturas cuando se fusionan para este cometido.

En lo referido al efecto “de bloques” y “borroso”, se podrían establecer distintas formas de proceder con la región solapada, que no sean tan básicas como una normalización al uso. La aplicación de una media ponderada por el inverso de la distancia al centro del parche, podría quizá suavizar todavía más ambos efectos. Sobre todo ayudaría con el efecto “de bloques”, porque introduciría una progresión o profundidad en el área solapada, en función de la localización del píxel.

Por último, sería interesante implementar síntesis más realistas de las imágenes ponderadas a partir de los mapas, para así predecir correctamente de forma combinada tanto la imagen post-T1w sintética, como los mapas post-T1 y post-T2.

Bibliografía

[1] Antonio Omuro and Lisa M. DeAngelis. Glioblastoma and Other Malignant Gliomas: A Clinical Review. *JAMA*, 310(17):1842–1850, 11 2013.

[2] Linda M. Wang, Zachary K. Englander, Michael L. Miller, and Jeffrey N. Bruce. *Malignant Glioma*, pages 1–30. Springer International Publishing, Cham, 2023.

[3] Jigisha P Thakkar, Pier Paolo Peruzzi, and Vikram C Prabhu, 2021. Website. Último acceso ago. 2024. `https://www.aans.org/patients/conditions-treatments/glioblastoma-multiforme/`.

[4] Elisa Moya Sáez. *Enhancing brain tumor diagnosis with Synthetic MRI*. PhD thesis, Universidad de Valladolid, 2024.

[5] Vikas Gulani and Nicole Seiberlich. Quantitative MRI: Rationale and challenges. In Nicole Seiberlich, Vikas Gulani, Fernando Calamante, Adrienne Campbell-Washburn, Mariya Doneva, Houchun Harry Hu, and Steven Sourbron, editors, *Quantitative Magnetic Resonance Imaging*, volume 1 of *Advances in Magnetic Resonance Technology and Applications*, pages xxxvii–li. Academic Press, 2020.

[6] Amaresha Shridhar Konar, Ramesh Paudyal, Akash Deelip Shah, Maggie Fung, Suchandrima Banerjee, Abhay Dave, Nancy Lee, Vaios Hatzoglou, and Amita Shukla-Dave. Qualitative and quantitative performance of magnetic resonance image compilation (MAGiC) method: An exploratory analysis for head and neck imaging. *Cancers*, 14(15), 2022.

[7] Benjamin M Ellingson, Martin Bendszus, Jerrold Boxerman, Daniel Barboriak, Bradley J Erickson, Marion Smits, Sarah J Nelson, Elizabeth Gerstner, Brian Alexander, Gregory Goldmacher, et al. Consensus recommendations for a standardized brain tumor imaging protocol in clinical trials. *Neuro-oncology*, 17(9):1188–1198, 2015.

[8] SC Thust, S Heiland, A Falini, Hans R Jäger, AD Waldman, PC Sundgren, Claudia Godi, VK Katsaros, A Ramos, N Bargallo, et al. Glioma imaging in europe: A survey of 220 centres and recommendations for best clinical practice. *European radiology*, 28:3306–3317, 2018.

[9] Ashkan Heshmatzadeh Behzadi, Yize Zhao, Zerwa Farooq, and Martin R Prince. Immediate allergic reactions to gadolinium-based contrast agents: a systematic review and meta-analysis. *Radiology*, 286(2):471–482, 2018.

[10] Kerry A Layne, Paul I Dargan, John RH Archer, and David M Wood. Gadolinium deposition and the potential for toxicological sequelae–a literature review of issues surrounding gadolinium-based contrast agents. *British journal of clinical pharmacology*, 84(11):2522–2534, 2018.

[11] Justyna Rogowska, Ewa Olkowska, Wojciech Ratajczyk, and Lidia Wolska. Gadolinium as a new emerging contaminant of aquatic environments. *Environmental toxicology and chemistry*, 37(6):1523–1534, 2018.

[12] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31:685–695, 2021.

[13] Pycharm - Python IDE for data science and web development. Website. Último acceso ago. 2024. `https://www.jetbrains.com/pycharm/`.

[14] The MRtrix3 - Advanced tools for the analysis of diffusion MRI data. Website. Último acceso ago. 2024. `https://www.mrtrix.org/`.

[15] Python - The official home of the Python Programming Language. Website. Último acceso ago. 2024. `https://www.python.org/`.

[16] LaTeX. Website. Último acceso ago. 2024. `https://www.latex-project.org/`.

[17] Overleaf - Online LaTeX editor. Website. Último acceso ago. 2024. `https://es.overleaf.com/`.

[18] The MathWorks Inc. Matlab version: 9.11.0 (r2021b), 2022.

[19] Robert H. Caverly. MRI fundamentals: RF aspects of magnetic resonance imaging (MRI). *IEEE Microwave Magazine*, 16(6):20–33, 2015.

[20] WR Hendee and CJ Morgan. Magnetic resonance imaging. Part I–physical principles. *The Western journal of medicine*, 141(4):491—500, October 1984.

[21] Victor Rakesh Lazar. *Quantification of bone using a 3.0 tesla clinical magnetic resonance scanner*. PhD thesis, University of Hull, 2011.

[22] AO Rodriguez. Principles of magnetic resonance imaging. *Revista mexicana de física*, 50(3):272–286, 2004.

[23] JA Soto, T Córdova, M Sosa, and S Jerez. Quantum-mechanical aspects of magnetic resonance imaging. *Revista mexicana de física E*, 63(1):48–55, 2017.

[24] Harish A. Sharma and Jim Lagopoulos. MRI physics: pulse sequences. *Acta Neuropsychiatrica*, 22(2):90–92, 2010.

[25] Carlos FGC Geraldes and Sophie Laurent. Classification and basic properties of contrast agents for magnetic resonance imaging. *Contrast media & molecular imaging*, 4(1):1–23, 2009.

[26] Costas D Arvanitis, Gino B Ferraro, and Rakesh K Jain. The blood–brain barrier and blood–tumour barrier in brain tumours and metastases. *Nature Reviews Cancer*, 20(1):26–41, 2020.

[27] Osval Antonio Montesinos López, Abelardo Montesinos López, and Jose Crossa. *Fundamentals of Artificial Neural Networks and Deep Learning*, pages 379–597. Springer International Publishing, Cham, 2022.

[28] Jahaziel Ponce. Funciones de activación y cómo puedes crear la tuya usando Python, R y TensorFlow, 2020.

[29] Phung and Rhee. A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. *Applied Sciences*, 9:4500, 10 2019.

[30] GeeksforGeeks. CNN - Introduction to Pooling Layer. `https://www.geeksforgeeks.org/cnn-introduction-to-pooling-layer/`, 2024.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted*

intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, pages 234–241. Springer, 2015.

[32] GeeksforGeeks. U-Net architecture explained. `https://www.geeksforgeeks.org/u-net-architecture-explained/`, 2021.

[33] Yufei Li, Ning Han, Yanyan Qin, et al. Trans-cGAN: redes generativas adversarias basadas en transformadores-unet para la síntesis de imágenes de resonancia magnética multimodal. *Statistics and Computing*, 33(113), 2023.

[34] Reza Kalantar, Christina Messiou, Jessica M Winfield, Alexandra Renn, Arash Latifoltojar, Kate Downey, Aslam Sohaib, Susan Lalondrelle, Dow-Mu Koh, and Matthew D Blackledge. CT-based pelvic T1-weighted MR image synthesis using UNet, UNet++ and cycle-consistent generative adversarial network (Cycle-GAN). *Frontiers in Oncology*, 11:665807, 2021.

[35] Pavlo Radiuk. Applying 3D U-Net architecture to the task of multi-organ segmentation in computed tomography. *Applied Computer Systems*, 25(1):43–50, 2020.

[36] Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. ResViT: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10):2598–2614, 2022.

[37] Nicolae-Cătălin Ristea, Andreea-Iuliana Miron, Olivian Savencu, Mariana-Iuliana Georgescu, Nicolae Verga, Fahad Shahbaz Khan, and Radu Tudor Ionescu. CyTran: A cycle-consistent transformer with multi-level consistency for non-contrast to contrast CT translation. *Neurocomputing*, 538:126211, 2023.

[38] Alan C Evans, Andrew L Janke, D Louis Collins, and Sylvain Baillet. Brain templates and atlases. *Neuroimage*, 62(2):911–922, 2012.

[39] Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, et al. Automated brain extraction of multisequence MRI using artificial neural networks. *Human brain mapping*, 40(17):4952–4964, 2019.

[40] Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841, 2002.

[41] Philipp Kickingereder, Fabian Isensee, Irada Tursunova, Jens Petersen, Ulf Neuberger, David Bonekamp, Gianluca Brugnara, Marianne Schell, Tobias Kessler, Martha Foltyn, et al. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *The Lancet Oncology*, 20(5):728–740, 2019.

[42] Mario Campos Soberanis. Introducción al deep learning I: Funciones de activación. `https://medium.com/soldai/introducciÃşn-al-deep-learning-i-funciones-de-activaciÃşn-b3eed1411b20`, 2023.

[43] Pranjal Datta. All about Structural Similarity Index (SSIM) - theory & code in PyTorch. `https://medium.com/srm-mic/all-about-structural-similarity-index-ssim-theory-code-in-pytorch-6551b455541e`, 2023.

[44] Dimple Sethi, Sourabh Bharti, and Chandra Prakash. A comprehensive survey on gait analysis: History, parameters, approaches, pose estimation, and future work. *Artificial Intelligence in Medicine*, 129:102314, 2022.

Apéndice A

Parámetros de adquisición de la base de datos

Tabla A.1: Parámetros de adquisición de la base de datos *GLIOMA*

	GLIOMA				
	T1w (IR-GRE)	**T2w** (TSE)	**T2w-FLAIR** (IR-TSE)	**post-T1w** (IR-GRE)	**MAGiC** (MDME)
TE (ms)	3.3	97	89	3.3	6114
TR (ms)	7.9	9837	5000	7.9	15.7
TI (ms)	450	-	1588	450	11
α (°)	12	90	90	12	90
# Cortes (orientación)	352 (axial)	49 (axial)	224 (sagital)	352 (axial)	49 (axial)
Grosor de corte (mm)	1.2	3.0	1.6	1.0	3.0
Resolución (mm^2)	1.0×1.0	0.6×0.6	1.11×1.11	1.0×1.0	1.0×1.0
FOV (mm)	240×240	233×233	246×246	240×240	240×240
Tiempo de adquisición (min)	$\sim$ 5:00	$\sim$ 4:00	$\sim$ 4:00	$\sim$ 5:00	$\sim$ 5:00