



UNIVERSITAT OBERTA DE CATALUNYA (UOC)

MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS (*Data Science*)

TRABAJO FINAL DE MÁSTER

ÁREA: COMPUTER VISION AND DEEP LEARNING APPLIED TO MEDICAL
IMAGING

Detección de lesiones nuevas o cambiantes en EM

Autor: Sofía Ramírez López

Tutor: Eloy Martínez de las Heras

Profesor: Laia Subirats Maté

Madrid, 9 de enero de 2026

Créditos/Copyright



Esta obra está sujeta bajo la licencia de Creative Commons [Attribution-NonCommercial-ShareAlike 4.0 International](#).

FICHA DEL TRABAJO FINAL

Título del trabajo:	Detección de lesiones nuevas o cambiantes en EM
Nombre del autor:	Sofía Ramírez López
Nombre del colaborador/a docente:	Eloy Martínez de las Heras
Nombre del PRA:	Laia Subirats Maté
Fecha de entrega (mm/aaaa):	01/2026
Titulación o programa:	Máster Universitario en Ciencia de Datos
Área del Trabajo Final:	Computer vision and deep learning applied to medical imaging
Idioma del trabajo:	Español
Palabras clave	Esclerosis múltiple, MRI, nnU-Net, PH
Repositorio GitHub:	https://github.com/sramirezlo/EM_seg_long

Nada en la vida debe ser temido, sólo comprendido.

Ahora es el momento de comprender más, para temer menos.

Marie Curie

Agradecimientos

En primer lugar me gustaría agradecer la labor que realizan los profesionales médicos e investigadores, que dedican hasta su alma para buscar un tratamiento para la esclerosis múltiple.

A los radiólogos, gracias a las segmentaciones manuales de las lesiones que han realizado, este proyecto ha sido posible.

Por otro lado, también agradezco que las personas afectadas con esclerosis múltiple hayan dado el consentimiento para que sus imágenes tomadas por resonancia magnética puedan usarse con el fin de investigar sobre modelos que sean capaces de detectar las lesiones con alta precisión, permitiendo que en un futuro se pueda hacer un seguimiento de la enfermedad con menos tiempos de espera.

Agradezco a mi familia que me haya estado aguantando todos estos meses sin llegar a odiarme.

Mónica, me sacaste de la desesperación y me has estado dando muy buenos consejos y apoyando en todo momento, muchas gracias.

Muchas gracias a ti también Anto, han sido meses difíciles, pero siempre has estado ahí para escucharme y apoyarme.

También agradezco la labor de mi tutor, Eloy, perdón por mandar correos a horas intempestivas. Muchas gracias por tus consejos y las indicaciones que me has ido dando.

Resumen / Abstract

Resumen

La esclerosis múltiple (EM) es una enfermedad crónica autoinmune que ataca la mielina del sistema nervioso central, formado por el encéfalo y la médula espinal. La mielina es el protector de las fibras nerviosas (axones). Cuando esta capa protectora desaparece, el axón queda expuesto al ataque también, produciéndose una pérdida de neuronas.

Generalmente se dan los primeros síntomas entre los 20 y 40 años y suele diagnosticarse a través de imágenes de resonancias magnéticas (RM). Actualmente no existe cura para esta enfermedad, pero los tratamientos pueden prevenir nuevos brotes y paliar síntomas. Es muy importante el seguimiento de la enfermedad a través de la realización de nuevas imágenes por RM, que permiten analizar el avance de la enfermedad y la eficacia del tratamiento en cada persona.

Este trabajo desarrolla un modelo de aprendizaje profundo basado en la red convolucional nnU-Net (framework más flexible y competitivo en segmentación médica desde 2020) que, a partir de imágenes longitudinales cerebrales por RM (secuencia FLAIR) tomadas en el Hospital Clínic de Barcelona a 349 personas diagnosticadas con EM, detecta nuevas lesiones o cambios en lesiones comparando imágenes basales con imágenes de seguimiento, asignando la etiqueta “0” a las zonas sin lesiones, “1” a las lesiones sin cambios y “2” a las lesiones nuevas.

Para reducir los falsos positivos en las predicciones, preservando la sensibilidad, se usan técnicas de postprocesamiento basadas en homología persistente (PH). El modelo con PH obtuvo un Dice de 0.8629 para la detección de lesiones nuevas en el test-split del conjunto ImaginEM y 0.3834 en el dataset externo MSSEG2, mejorando la precisión diagnóstica y convirtiendo el modelo en una herramienta de apoyo para el especialista en la segmentación.

Palabras clave: Esclerosis múltiple, nnU-Net, Homología persistente, Aprendizaje profundo, Segmentación médica, Resonancia magnética.

Abstract

Multiple sclerosis (MS) is a chronic autoimmune disease that attacks the myelin of the central nervous system, which is formed by the brain and the spinal cord. Myelin acts as a protective layer for nerve fibers (axons). When this protective layer disappears, the axon becomes exposed to damage as well, leading to the loss of neurons.

The first symptoms usually appear between the ages of 20 and 40, and the disease is generally diagnosed through magnetic resonance imaging (MRI). Currently, there is no cure for multiple sclerosis, but treatments can prevent new relapses and alleviate symptoms. It is very important to monitor the disease by performing new MRI scans, which allow us to analyze its progression and evaluate the effectiveness of treatment for each patient.

This work develops a deep learning model based on the convolutional network nnU-Net (the most flexible and competitive framework for medical image segmentation since 2020). Using longitudinal brain MRI images (FLAIR sequences) from 349 people diagnosed with MS at Hospital Clínic de Barcelona to detect new lesions or changes in existing ones by comparing baseline and follow-up images. The model assigns the label “0” to areas without lesions, “1” to unchanged lesions, and “2” to new lesions.

To reduce false positives in predictions while preserving sensitivity, post-processing techniques based on persistent homology (PH) are used. The PH model achieved a Dice score of 0.8629 for new lesion detection on the ImaginEM test-split and 0.3834 on the external MSSEG2 dataset, improving diagnostic precision and turning the model into a support tool for specialists in segmentation.

Keywords: Multiple sclerosis, nnU-Net, Persistent homology, Deep learning, Medical segmentation, Magnetic resonance imaging.

Índice general

Resumen / Abstract	ix
Índice	xii
Lista de Figuras	xv
Lista de Tablas	xvii
1. Introducción	3
1. Contexto y motivación	3
2. Objetivos	7
3. Sostenibilidad, diversidad y desafíos ético/sociales	7
4. Enfoque y metodología	8
5. Planificación	10
6. Resumen de los productos del proyecto	11
7. Breve descripción de los demás capítulos del informe	12
2. Estado del arte	13
1. Contexto histórico y primeros pasos	13
2. Avances en Deep Learning y técnicas actuales	14
3. Nuevas líneas de investigación	16
3. Métodos y recursos	19
1. Conjunto de datos ImaginEM	19
1.1. Preprocesamiento imágenes RM ImaginEM	20
1.2. Estudio estadístico del conjunto de datos ImaginEM	21
2. Conjunto de datos MSSEG2	22
2.1. Estudio estadístico del conjunto de datos MSSEG2	22
3. Definiciones previas	23
3.1. Métricas	27

4.	Procedimiento	27
4.1.	Estrategias entrenamiento	28
4. Resultados		35
1.	Evaluación test-split	36
1.1.	Voxel-wise ImaginEM	36
1.2.	Lesion-wise ImaginEM	36
1.3.	ID-wise ImaginEM	37
2.	Evaluación MSSEG2	37
2.1.	Voxel-wise MSSEG2	37
2.2.	Lesion-wise MSSEG2	38
2.3.	ID-wise MSSEG2	38
3.	Postprocesado	38
3.1.	Estrategias postprocesado	39
3.1.1.	Postprocesado por volumen de lesiones	39
3.1.1.1.	Voxel-wise ImaginEM	39
3.1.1.2.	Lesion-wise ImaginEM	40
3.1.1.3.	ID-wise ImaginEM	40
3.1.1.4.	Voxel-wise MSSEG2	40
3.1.1.5.	Lesion-wise MSSEG2	41
3.1.1.6.	ID-wise MSSEG2	41
3.1.2.	Postprocesado por homología persistente	42
3.1.2.1.	Voxel-wise ImaginEM	45
3.1.2.2.	Lesion-wise ImaginEM	46
3.1.2.3.	ID-wise ImaginEM	46
3.1.2.4.	Voxel-wise MSSEG2	46
3.1.2.5.	Lesion-wise MSSEG2	46
3.1.2.6.	ID-wise MSSEG2	46
3.1.3.	Postprocesado usando PH+volumen (PH+P3)	47
3.1.4.	Postprocesado usando volumen+PH (P3+PH)	47
3.1.4.1.	Voxel-wise ImaginEM	47
3.1.4.2.	Lesion-wise ImaginEM	47
3.1.4.3.	ID-wise ImaginEM	47
3.1.4.4.	Voxel-wise MSSEG2	48
3.1.4.5.	Lesion-wise MSSEG2	48
3.1.4.6.	ID-wise MSSEG2	48
4.	Conclusiones análisis cuantitativo de los distintos postprocesamientos	49

5.	Análisis cualitativo de los postprocesados en ImaginEM	50
6.	Análisis cualitativo del postprocesado en MSSEG2	54
5.	Conclusiones y trabajo futuro	57
1.	Conclusiones finales	57
2.	Limitaciones	58
3.	Áreas de mejora	59
4.	Próximos pasos	60

Índice de figuras

1.1.	FLAIR baseline vs FLAIR followup vs GT lesión basal.	4
1.2.	FLAIR baseline vs FLAIR followup vs GT lesión nueva.	5
1.3.	Máscara de segmentación manual de una lesión estable en eje axial, coronal y sagital.	5
1.4.	T1-w vs T2-w vs FLAIR [21]	6
1.5.	Planificación temporal del TFM: Detección de lesiones nuevas o cambiantes en EM.	10
1.6.	Pipeline del proyecto.	11
2.1.	Arquitectura U-Net para segmentación de lesiones [19]	14
3.1.	Máscara de segmentación manual. Las lesiones rodeadas de verde corresponden con lesiones estables, las rosas con lesiones nuevas.	20
3.2.	Gráficas estudio estadístico del conjunto ImaginEM	22
3.3.	Gráficas estudio estadístico del conjunto MSSEG2	23
3.4.	Representación píxel vs representación voxel	24
3.5.	Representación 3D de lesiones en voxels	25
3.6.	Voxeles de lesiones	26
3.7.	Entrenamiento modelo HOLDOUT con 20 épocas.	30
3.8.	Entrenamiento modelo HOLDOUT con 50 épocas.	31
3.9.	Entrenamiento modelo HOLDOUT con 100 épocas.	32
3.10.	Entrenamiento modelo HOLDOUT con 250 épocas.	33
4.1.	Gráficos violín estadísticos PH.	44
4.2.	Segmentación manual vs segmentación modelo. En este caso las predicciones de lesiones nuevas son buenas.	51
4.3.	Segmentación 100 VOL P3 vs segmentación 100 PH K3.	51
4.4.	Segmentación manual vs segmentación modelo. En este caso las predicciones de lesiones nuevas son malas.	52

4.5. Segmentación manual vs segmentación modelo. En este caso las predicciones de lesiones estables son buenas.	52
4.6. Segmentación manual vs segmentación modelo. En este caso las predicciones de lesiones estables son malas.	53
4.7. Comparativa TP, FP, FN en una predicción de lesión estable en ImaginEM.	54
4.8. Comparativa TP, FP, FN en una predicción de lesión nueva en ImaginEM.	54
4.9. Evolución de la lesión en MSSEG2 y segmentación 100 PH K3.	55
4.10. Evolución de la lesión en MSSEG2 y segmentación 100 PH K3.	55
4.11. Evolución de la lesión en MSSEG2 y segmentación 100 PH K3.	55
4.12. Evolución de la lesión en MSSEG2 y segmentación 100 PH K3.	56

Índice de cuadros

2.1. Comparativa de arquitecturas para segmentación de lesiones.	15
3.1. Resumen estadístico del conjunto de datos ImaginEM.	21
3.2. Resumen estadístico del conjunto de datos MSSEG2.	23
3.3. Espaciado voxel y dimensiones en ImaginEM.	24
4.1. Resultados voxel-wise para la lesión nueva en test-split ImaginEM.	36
4.2. Métricas voxel-wise para la lesión estable y promedios en test-split ImaginEM. .	36
4.3. Resultados lesion-wise en test-split ImaginEM.	36
4.4. Resultados ID-wise en test-split ImaginEM.	37
4.5. Resultados voxel-wise para la lesión nueva en MSSEG2.	37
4.6. Métricas voxel-wise para voxels sin lesión nueva y promedios en MSSEG2. . .	37
4.7. Resultados lesion-wise en MSSEG2.	38
4.8. Resultados ID-wise en MSSEG2.	38
4.9. Resultados voxel-wise para la lesión nueva en ImaginEM con postprocesado volumen.	39
4.10. Resultados voxel-wise para la lesión estable y promedios en ImaginEM con postprocesado volumen.	40
4.11. Resultados lesion-wise en ImaginEM para el modelo de 100 épocas con postprocesado volumen.	40
4.12. Resultados ID-wise en ImaginEM para el modelo de 100 épocas con postprocesado volumen.	40
4.13. Resultados voxel-wise para la lesión nueva en MSSEG2 modelo de 100 épocas con postprocesado volumen.	41
4.14. Resultados voxel-wise para voxels sin lesión nueva y promedios en MSSEG2 modelo de 100 épocas con postprocesado volumen.	41
4.15. Resultados lesion-wise en MSSEG2 para el modelo de 100 épocas con postprocesado volumen.	41

4.16. Resultados ID-wise en MSSEG2 para el modelo de 100 épocas con postprocesado volumen.	41
4.17. Estadísticos descriptivos de <i>n_bars_h0_fl</i> en FLAIR followup para falsos positivos (FP) y verdaderos positivos (TP).	43
4.18. Estadísticos descriptivos de <i>sum_life_h0_fl</i> en FLAIR followup para falsos positivos (FP) y verdaderos positivos (TP).	44
4.19. Fracción de verdaderos positivos (TP_keep) y falsos positivos (FP_keep) retenidos para distintos valores de <i>k</i> y del umbral τ	45
4.20. Resultados voxel-wise para la lesión nueva en ImaginEM (modelo 100_PH_K3).	45
4.21. Resultados voxel-wise para la lesión estable y promedios en ImaginEM (modelo 100_PH_K3).	45
4.22. Resultados lesion-wise en ImaginEM para el modelo 100_PH_K3.	46
4.23. Resultados ID-wise en ImaginEM para el modelo 100_PH_K3.	46
4.24. Resultados voxel-wise para la lesión nueva en MSSEG2 (modelo 100_PH_K3).	46
4.25. Resultados voxel-wise para voxels sin lesión nueva y promedios en MSSEG2 (modelo 100_PH_K3).	46
4.26. Resultados lesion-wise en MSSEG2 para el modelo 100_PH_K3.	46
4.27. Resultados ID-wise en MSSEG2 para el modelo 100_PH_K3.	46
4.28. Resultados voxel-wise para la lesión nueva en ImaginEM (modelos 100_K3_P3 y 100_P3_K3).	47
4.29. Resultados voxel-wise para la lesión estable y promedios en ImaginEM (modelos 100_K3_P3 y 100_P3_K3).	47
4.30. Resultados lesion-wise en ImaginEM para los modelos 100_K3_P3 y 100_P3_K3.	47
4.31. Resultados ID-wise en ImaginEM para los modelos 100_K3_P3 y 100_P3_K3.	48
4.32. Resultados voxel-wise para la lesión nueva en MSSEG2 (modelos 100_K3_P3 y 100_P3_K3).	48
4.33. Resultados voxel-wise para voxels sin lesión nueva y promedios en MSSEG2 (modelos 100_K3_P3 y 100_P3_K3).	48
4.34. Resultados lesion-wise en MSSEG2 para los modelos 100_K3_P3 y 100_P3_K3.	48
4.35. Resultados ID-wise en MSSEG2 para los modelos 100_K3_P3 y 100_P3_K3.	48
4.36. Métricas voxel-wise por clase para el caso ImaginEM_MSVIS_095_03 con el modelo 100_PH_K3.	51
4.37. Métricas voxel-wise por clase para el caso ImaginEM_MSVIS_072_05 con el modelo 100_PH_K3.	52
4.38. Métricas voxel-wise por clase para el caso ImaginEM_MSVIS_089_02 con el modelo 100_PH_K3.	53

4.39. Métricas voxel-wise por clase para el caso ImaginEM_MSVIS_088_05 con el modelo 100_PH_K3. 53

Capítulo 1

Introducción

La esclerosis múltiple es una enfermedad crónica autoinmune que ataca la mielina del sistema nervioso central, formado por el encéfalo y la médula espinal. La mielina es el protector de las fibras nerviosas (axones). Cuando esta capa protectora desaparece, el axon queda expuesto al ataque también, produciéndose una pérdida de neuronas. Actualmente no existe cura para esta enfermedad, pero los tratamientos pueden prevenir nuevos brotes y paliar síntomas. Es muy importante el seguimiento de la enfermedad a través de la realización de nuevas imágenes por resonancias magnéticas (RM) y de su estudio, que permiten analizar el avance de la enfermedad y la eficacia del tratamiento en cada persona. Años atrás únicamente podían estudiar las imágenes los radiólogos, lo que implicaba posibles fallos humanos o sesgos dependiendo de la experiencia de cada especialista. Pero, con el avance de la inteligencia artificial, en concreto con las redes neuronales y la visión por computador, se han ido desarrollando modelos de aprendizaje profundo capaces de detectar estas lesiones, ahorrando tiempo de espera para las personas, y liberando la carga laboral de los profesionales radiólogos.

1. Contexto y motivación

Con el fin de mejorar estos modelos automáticos de detección, actualmente hay muchos investigadores que desarrollan modelos de aprendizaje profundo basados en redes convolucionales para detectar nuevas lesiones o ver la evolución de las lesiones en personas con esclerosis múltiple y acelerar su diagnóstico, puesto que a día de hoy es una enfermedad que no tiene cura, y su detección temprana puede hacer que no ocurran brotes que afecten gravemente a las personas que tienen esta enfermedad.

El punto de partida para ello, es la recopilación de imágenes longitudinales de RM para cada una de las personas afectadas. Las imágenes longitudinales de un sujeto consisten en imágenes de la misma zona afectada a estudiar pero tomadas en distintos momentos temporales. Por

ejemplo, la imagen de la primera RM en la que se le diagnosticó la enfermedad a la persona (imagen basal) y la imagen por RM un año después en esa misma zona (imagen de seguimiento). La confrontación longitudinal de imágenes por RM de la misma zona permite identificar cambios estructurales con el fin de detectar nuevas lesiones o cambios en lesiones existentes en caso de que las haya, en cada uno de los sujetos. A continuación se muestran dos ejemplos ilustrativos.

En el siguiente panel se muestra la evolución de una lesión para un ID del conjunto de datos ImaginEM. La imagen FLAIR baseline corresponde a la imagen tomada en el momento t_1 , y la imagen FLAIR followup corresponde a la imagen tomada en el momento t_2 , con $1 \leq t_2 - t_1 \leq 3$ años. En la tercera imagen se representa en verde la segmentación manual realizada por el especialista. En este caso se observa que entre t_1 y t_2 no ha habido cambios en la lesión, por lo que se etiqueta como lesión estable.

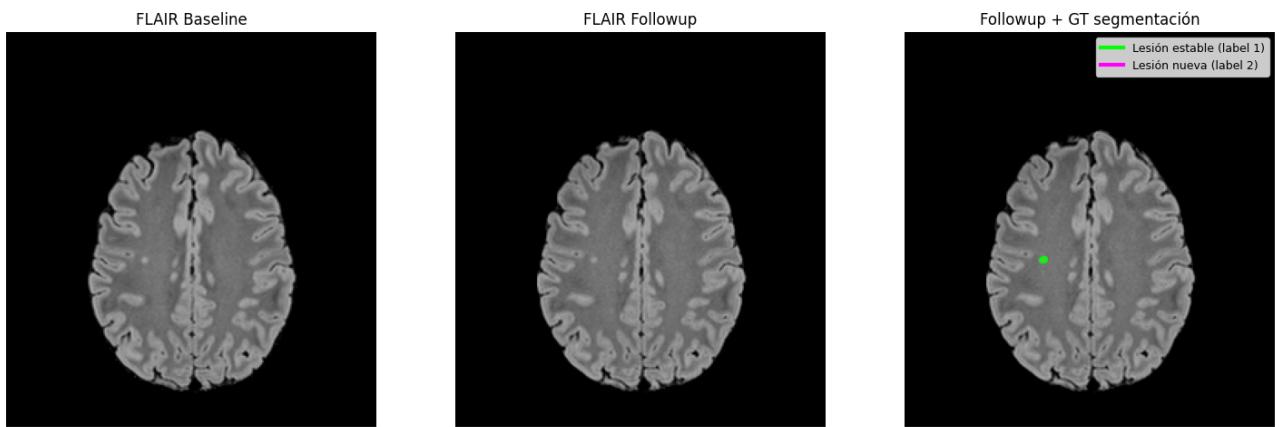


Figura 1.1: FLAIR baseline vs FLAIR followup vs GT lesión basal.

En cambio, en la siguiente imagen se observa cómo entre la imagen tomada en el momento t_1 (FLAIR baseline), y t_2 (FLAIR followup) aparece una lesión. Esto se ve reflejado en la tercera imagen, donde en la segmentación manual aparece esta lesión representada en violeta, y etiquetada como nueva.

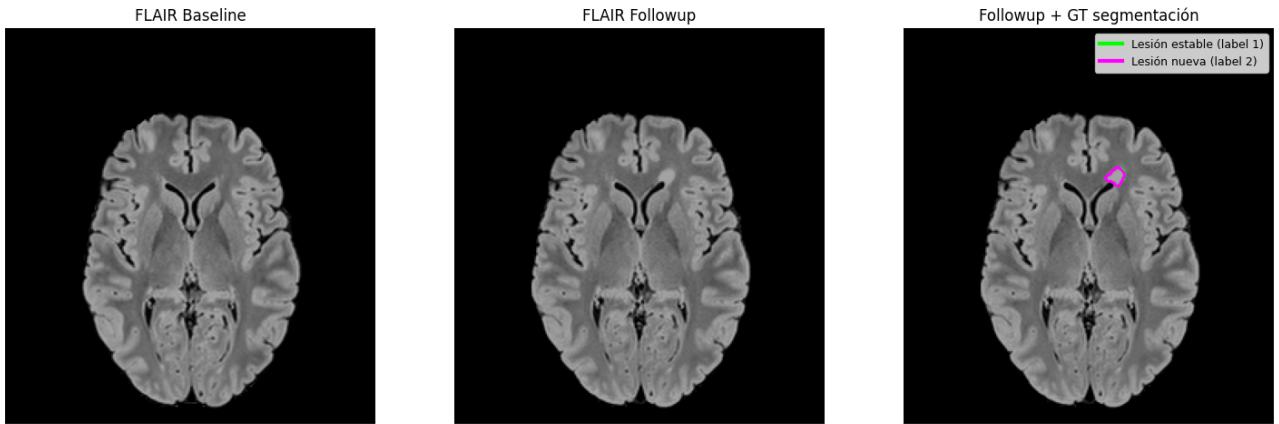


Figura 1.2: FLAIR baseline vs FLAIR followup vs GT lesión nueva.

Las imágenes anteriores están representadas en el eje axial, pero se pueden representar en otros dos ejes más. En el ejemplo siguiente se muestra para un ID distinto de ImaginEM la segmentación manual de una lesión basal (lesión estable) en los tres ejes (axial \equiv eje Z, coronal \equiv eje Y, sagital \equiv eje X)

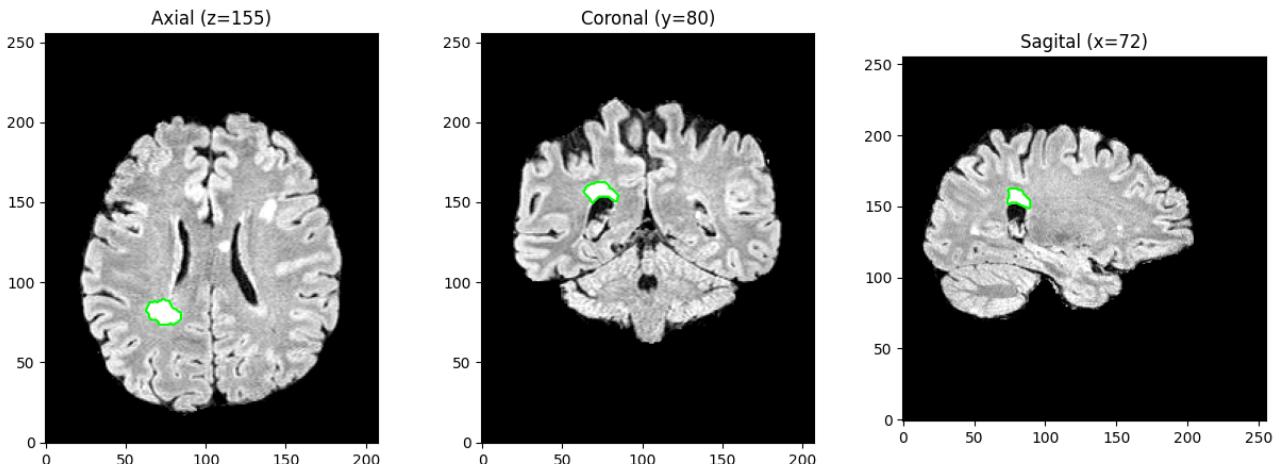


Figura 1.3: Máscara de segmentación manual de una lesión estable en eje axial, coronal y sagital.

Para obtener visualizaciones detalladas en las resonancias, se emplean distintas secuencias, cada una de ellas aporta diferentes contrastes entre los tejidos del cerebro. Las secuencias más utilizadas son: FLAIR (Fluid Attenuated Inversion Recovery) esta secuencia se encarga de suprimir las señales CSF (señal del líquido cefalorraquídeo, realzando las lesiones en la sustancia blanca), de forma que se pueda visualizar claramente las zonas del sistema nervioso central; T1-w (mide el tiempo en el que el átomo de hidrógeno recupera el equilibrio) y T2-w (mide el tiempo en el que el átomo pierde el equilibrio) que muestran la respuesta del núcleo de los átomos de hidrógeno cuando se realiza el escáner de la resonancia magnética; y PD (densidad

protónica), que muestra la cantidad de protones del átomo de hidrógeno en un tejido. Sobre todo, la secuencia que se usa principalmente es la FLAIR. T1-w, T2-w y PD pueden servir como apoyo para mejorar la eficacia del modelo desarrollado. En este trabajo se emplea la secuencia FLAIR. A continuación se muestra una imagen con las secuencias más usadas, junto con su máscara de segmentación manual (ground truth, GT). Se observa que es en la secuencia FLAIR donde mejor se distinguen las zonas que corresponden a lesiones.

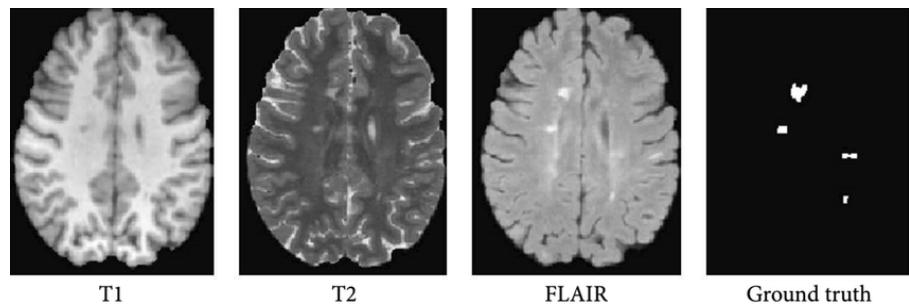


Figura 1.4: T1-w vs T2-w vs FLAIR [21]

Cabe destacar que los protocolos de obtención de imágenes por RM no están estandarizados globalmente, lo que hace que las propiedades de cada imagen no sean iguales si no se han tomado en la misma máquina con los mismos protocolos. Debido a ello, habría que estudiar su generalización entre distintas máquinas de resonancia y centros hospitalarios, para que estos modelos de detección de lesiones producidas por la esclerosis múltiple puedan aplicarse de forma general.

En este proyecto se van a emplear los conocimientos adquiridos en el Máster de Ciencia de Datos, con especialización en redes neuronales y aprendizaje profundo, para desarrollar un modelo de deep learning basado en la red nnU-Net que, a partir de imágenes longitudinales cerebrales de cada sujeto, obtenidas en el Hospital Clínic de Barcelona por resonancia magnética de una cohorte de 349 personas diagnosticados con esclerosis múltiple (EM), sea capaz de detectar de forma automática nuevas lesiones, o cambios en lesiones anteriores, en cada uno de los sujetos. De forma que, a las zonas sin lesiones le asigne la etiqueta 0, un 1 a las lesiones sin cambios y un 2 a las lesiones nuevas. También se emplearán técnicas de postprocesado basadas en la homología persistente, con el fin de reducir los falsos positivos de las predicciones, sin comprometer la sensibilidad. De este modo se reduce la carga de trabajo del especialista y se mejora la precisión diagnóstica. Con este proyecto se tiene la oportunidad también, de poder ayudar en el diagnóstico y seguimiento de la EM.

2. Objetivos

El principal objetivo es desarrollar un modelo basado en redes neuronales convolucionales (CNN) del tipo nnU-net, que detecte lesiones nuevas y cambiantes a partir de imágenes de RM en personas con EM. Se elige la red convolucional nnU-Net porque ha demostrado ser el framework más flexible y competitivo en segmentación de imágenes médicas desde 2020. Tiene una gran capacidad de generalización debido a que se adapta automáticamente la arquitectura y el preprocessamiento a los datos de cada problema. Esto ayuda mucho a la dificultad que se comentó anteriormente: la gran variabilidad de protocolos, secuencias, parámetros, resolución... en las máquinas de resonancias magnéticas de los distintos centros. Además, esta CNN incorpora procesos automáticos de preprocessamiento de imágenes como la normalización, el resampleo, la corrección de ruido u otras anomalías y aplica la técnica de augmentation para asegurar robustez frente a la variabilidad que se pueda dar. La configuración del entrenamiento también la realiza de forma automática, eligiendo los parámetros como el learning rate, el optimizador, número de épocas en función de las características del conjunto de datos que se usará para realizar el entrenamiento. Incorpora también la validación cruzada, para que el modelo sea capaz de aprender características globales y no sólo específicas. El resto de objetivos presentes son: comparar el rendimiento del modelo creado frente a segmentaciones realizadas de forma manual por radiólogos, y otros modelos de detección creados en competiciones como MSSEG2. Aplicar técnicas de postprocesamiento basadas en filtros de volumen de las lesiones y en homología persistente, con el fin de reducir los falsos positivos. Evaluar las métricas obtenidas por los modelos desarrollados, tales como la métrica Dice, F1, la precisión, sensibilidad... Documentar las limitaciones encontradas y discutir las posibles mejoras y los próximos pasos.

3. Sostenibilidad, diversidad y desafíos ético/sociales

A continuación se evalúa el impacto de este proyecto en los siguientes aspectos.

Sostenibilidad Para desarrollar este proyecto se ha requerido potencia de cómputo en la nube, en concreto, el uso de GPUs de Google Colab Pro+ para los entrenamientos del modelo y la ejecución del código. Según la infografía de NVIDIA, una GPU A100 consume aproximadamente 300W en cada instante en el que está en uso esta tarjeta gráfica. Para entrenar los distintos modelos y las respectivas pruebas, se ha requerido un total aproximado de 120 horas de cómputo de la A100, lo que hace, según la calculadora [Green Algorithms](#), un consumo aproximado de 41.17kWh, equivalente a 7.04kgCO₂e. Estas cifras se pueden traducir a un trayecto de 40.24km en coche. Por otro lado, los centros de datos de Google tiene un impacto medioambiental añadido, el uso del agua para refri-

gerar sus componentes. No hay cifras oficiales sobre la cantidad de agua que usa Google para la aceleración de procesos de aprendizaje automático mediante sus GPUs en la nube, pero en la web [Calculating the environmental footprint of AI at Google](#) estiman que en una consulta media en Gemini se usan alrededor de 0.26ml de agua por cada 0.24Wh de energía consumidos.

Comportamiento ético y responsabilidad social Las imágenes del dataset ImaginEM y las del conjunto de datos MSSEG2 han sido anonimizadas, de manera que no es posible revelar la identidad de cada ID. Además, entre la UOC (Universitat Oberta de Catalunya) y el IDIBAPS existe un comité de ética, que permite el uso de estas imágenes, siempre que estén anonimizadas.

El fin de este proyecto no es el de eliminar los puestos de trabajo de los radiólogos, sino el de conseguir generar un modelo que sea capaz de detectar de manera veraz lesiones nuevas o cambiantes, para que se convierta en una herramienta de apoyo para el radiólogo especialista.

Diversidad, género y derechos humanos Este proyecto únicamente usa imágenes anonimizadas que no poseen ningún tipo de información sobre sexo, origen étnico, religión, situación socioeconómica... por lo que no tiene impacto sobre la diversidad y el género. Lo que sí cabe mencionar es que las imágenes que se usan para entrenar el modelo han sido tomadas todas ellas en el Hospital Clínic de Barcelona, por lo que sí podría darse cierto sesgo, viéndose reflejado en que el modelo tuviera un menor rendimiento en imágenes procedentes de otros hospitales o incluso países, que hubieran sido tomadas con escáneres diferentes a los del Hospital Clínic de Barcelona. Por ello mismo el modelo entrenado se evalúa en datasets externos como MSSEG2. Todas las imágenes tienen el consentimiento de las personas que las cedieron y han sido anonimizadas, cumpliendo la GDPR (Reglamento General de Protección de Datos).

4. Enfoque y metodología

El proyecto se ha basado en el Método Científico siguiendo metodologías ágiles. Se han realizado muchos ensayos y se han obtenido muchos errores, y con ello se ha ido mejorando el desarrollo de los modelos, con el fin de seleccionar el que mejores métricas aportaba. El proceso de documentación ha sido exhaustivo, y se refleja en el [Capítulo 2](#). Las fases principales que tiene el proyecto son:

1. Documentación: proceso que ocurre de principio a fin del proyecto.

2. Obtención de datos: se accede a la base de datos donde se encuentran las imágenes tomadas por RM en el Hospital Clínic de Barcelona de las 349 personas con EM. Por cada persona habrá una imagen FLAIR baseline y una FLAIR followup.
3. Preprocesamiento de las imágenes: se aplica BET (Brain Extraction Tool) para coger de las imágenes únicamente el tejido cerebral, se usa la alineación espacial común del MNI (Montreal Neurological Institute), se corrigen inhomogeneidades usando el algoritmo N4, obteniendo imágenes estándares y consistentes.
4. Etiquetado: Se realiza por parte de los expertos del Hopital Clínic de Barcelona. Las etiquetas son las nombradas anteriormente, 0=“sin lesión”, 1=“lesión sin cambio”, 2=“lesión nueva”.
5. Desarrollo del modelo: Uso de la CNN nnU-Net como base. División del dataset ImaginEM en train/validation/test. Uso de técnicas de postprocesamiento basadas en la homología persistente para reducir los falsos positivos sin perder sensibilidad en el modelo.
6. Comparativa de resultados con el modelo MSSEG2. Evaluación de métricas.
7. Discusión de resultados: Análisis cuantitativo y cualitativo, análisis de limitaciones y propuestas de mejora.
8. Redacción y entrega de la memoria.
9. Defensa del proyecto.

5. Planificación

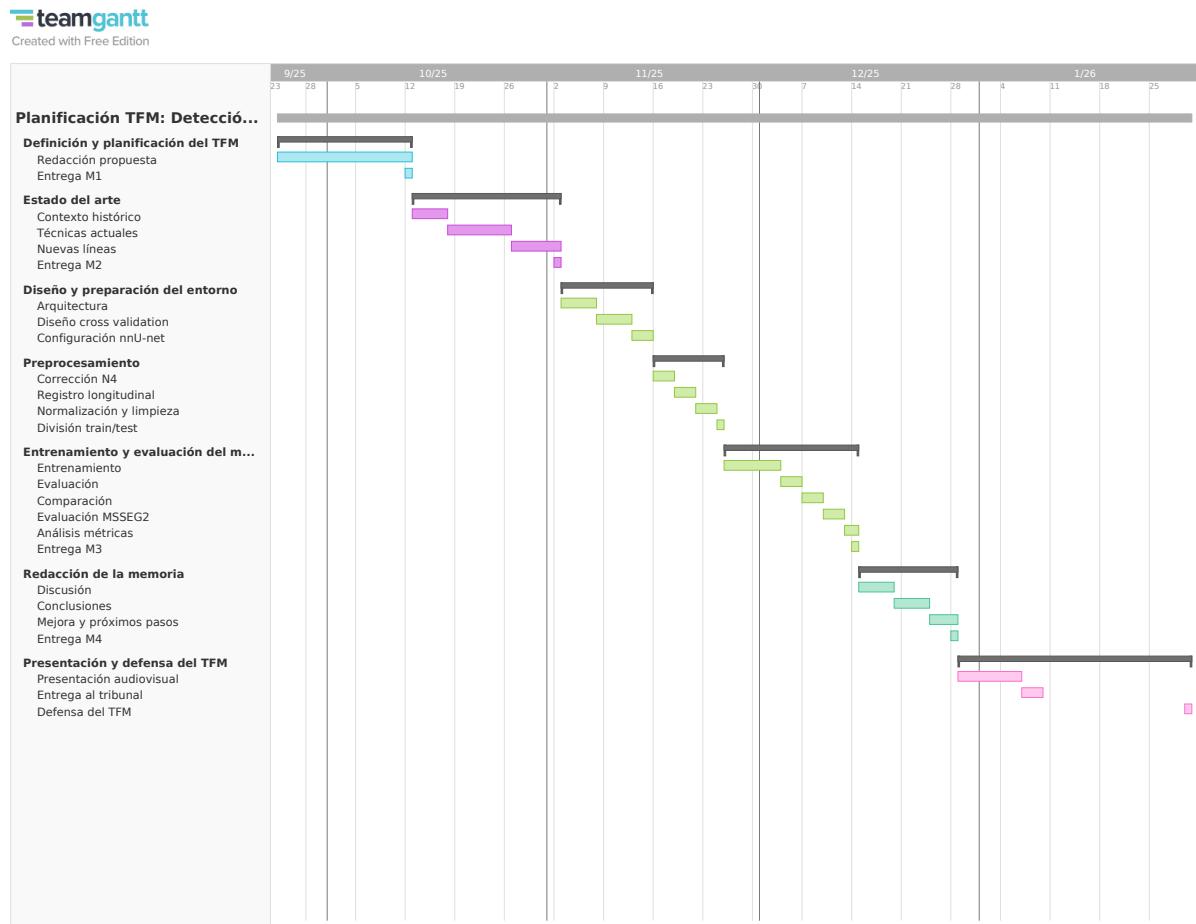


Figura 1.5: Planificación temporal del TFM: Detección de lesiones nuevas o cambiantes en EM.

6. Resumen de los productos del proyecto

En el proyecto se han desarrollado varios productos que todos ellos en conjunto conforman una solución para, a partir de dos imágenes, FLAIR baseline y FLAIR followup, junto con la máscara de segmentación manual (GT), desarrollar un modelo basado en la red neuronal convolucional nnU-Net v2 capaz de predecir lesiones estables (etiqueta 1) y lesiones nuevas (etiqueta 2) con alta precisión en el conjunto de datos ImaginEM. Para ello se han usado cuatro estrategias de entrenamiento (20 épocas, 50 épocas, 100 épocas y 250 épocas) y se ha escogido el modelo que daba mejor rendimiento en la evaluación en ImaginEM y MSSEG2, siendo el de 100 épocas el ganador. Después, con el fin de reducir el número de falsos positivos, se desarrollaron cuatro estrategias de postprocesamiento, basadas en filtros de volumen y de homología persistente, y la combinación de ambos tipos de filtros en distinto orden. Una vez que se evaluó el rendimiento de cada uno de estos cuatro postprocesamientos, el que mejores métricas aportaba en ImaginEM y MSSEG2 fue el postprocesado basado en la homología persistente. Por último, se discutieron los resultados obtenidos, se identificaron las áreas a mejorar y se definieron los próximos pasos para futuras investigaciones.

A continuación se muestra la implementación completa de este proyecto, desde el punto de partida (imágenes FLAIR), hasta el punto final, segmentaciones producidas por el modelo,

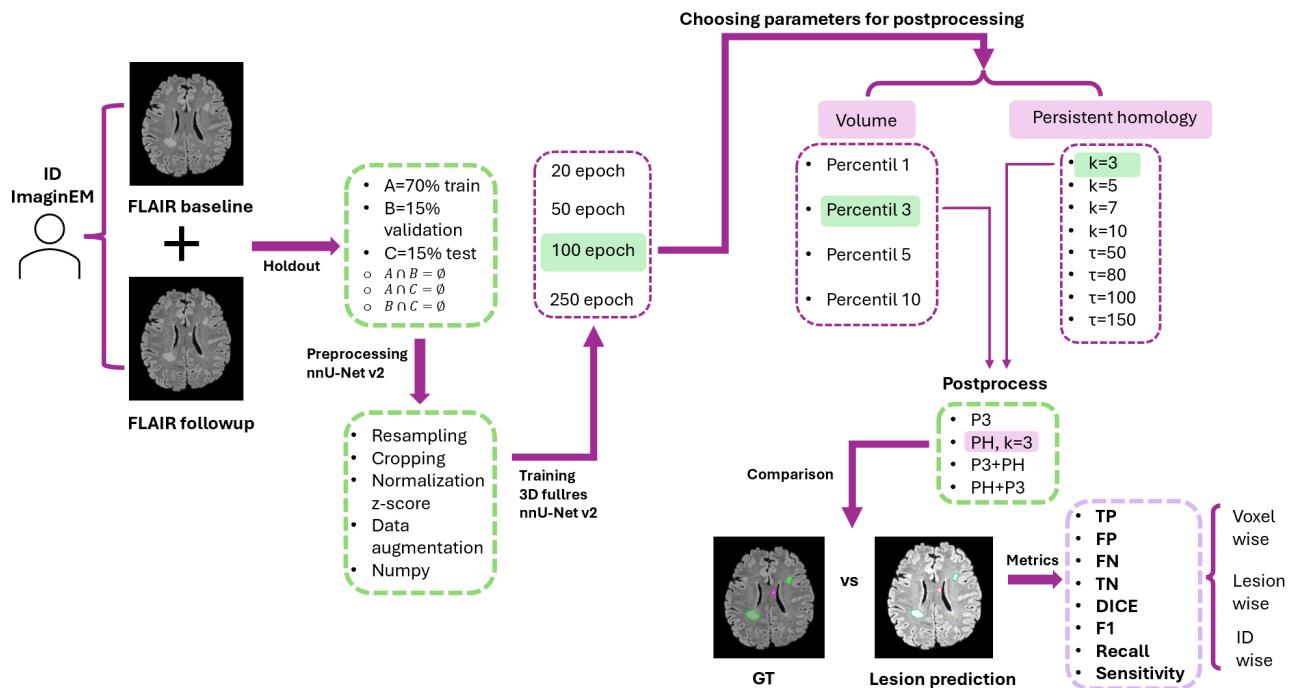


Figura 1.6: Pipeline del proyecto.

7. Breve descripción de los demás capítulos del informe

En [Capítulo 1](#), se contextualiza el tema sobre el que se centra el proyecto, se definen los objetivos, la motivación del proyecto y su planificación temporal.

En [Capítulo 2](#), se hace una revisión de las principales investigaciones que se han realizado para segmentar longitudinalmente, para cada uno de los IDs, las lesiones producidas por la EM. Destacando los modelos basados en la red nnU-Net e investigaciones recientes basadas en el Análisis de Datos Topológicos (TDA).

En [Capítulo 3](#), se realiza un análisis estadístico de los datasets ImaginEM y MSSEG2, se estudian las diferencias entre las imágenes de estos datasets, se introducen definiciones previas, y se describen las motivaciones y pasos de los distintos experimentos.

En [Capítulo 4](#), se exponen las métricas obtenidas de los modelos en el test-split de ImaginEM y en MSSEG2. También se realiza un análisis cualitativo. Para reducir FP se aplican cuatro técnicas de postprocesado y se evalúan los resultados obtenidos, mostrando cuál es la mejor técnica.

Por último, en [Capítulo 5](#), se recopilan las conclusiones obtenidas en el proyecto, se muestran las limitaciones detectadas, las áreas de mejora y se proponen las futuras líneas de investigación.

Capítulo 2

Estado del arte

La esclerosis múltiple (EM) es una enfermedad crónica autoinmune que ataca la mielina del sistema nervioso central, formado por el encéfalo y la médula espinal, provocando pérdida de neuronas, afectando de forma irreversible a las personas.

1. Contexto histórico y primeros pasos

Según el criterio de McDonald de 2017, las imágenes por resonancia magnética (RM) son una de las mejores herramientas para detectar lesiones cerebrales producidas por esta enfermedad. Retrocediendo en el tiempo, este cometido lo realizaban únicamente los radiólogos especialistas, sin apoyo alguno, lo que se traducía en largas esperas para analizar una por una cada imagen y, grandes cargas de trabajo para los especialistas. Con el avance de las nuevas tecnologías, varios grupos de investigadores trataron de realizar esta identificación de lesiones producidas por la EM, también conocido como segmentación de imágenes, mediante algoritmos de aprendizaje automático (machine learning, ML). La mayoría de estas investigaciones se han apoyado en secuencias T1, T2 y FLAIR de las RM. Estas primeras investigaciones basadas en ML empleaban SVM (Máquinas de Soporte de Vectores), random forests y clustering, y se apoyaban en características extraídas de forma manual como la textura o la forma para segmentar las lesiones.

Las limitaciones de estos procedimientos eran importantes, pues tenían una baja generalización ante conjuntos de datos muy variados, dependían de la calidad del etiquetado que realizaba el radiólogo y su capacidad para abordar imágenes longitudinales era mínima [17].

2. Avances en Deep Learning y técnicas actuales

Con la aparición de las redes neuronales convolucionales (CNN), y en particular la arquitectura U-net, se produjo una revolución en la segmentación de estas imágenes [19], ya que los resultados obtenidos eran precisos, y se podía capturar contexto espacial a diferentes escalas.

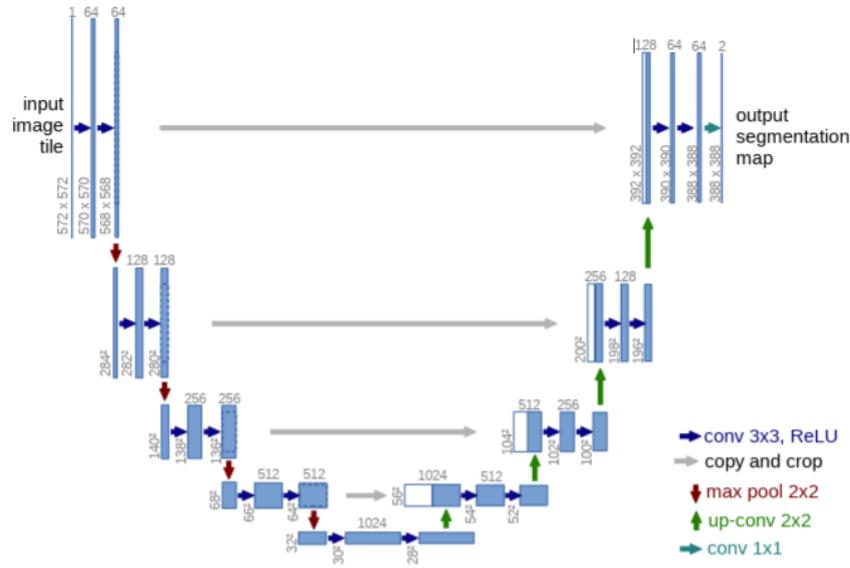


Figura 2.1: Arquitectura U-Net para segmentación de lesiones [19]

Uno de los primeros estudios basados en la red U-net fue el de [25], para segmentación de lesiones en secuencias FLAIR, obteniendo métricas comparables a los de la segmentación manual. A raíz de esto la comunidad científica se volcó en el estudio de estas arquitecturas, surgiendo nuevos modelos basados en arquitecturas que ofrecían mejores resultados. Como por ejemplo, su versión cúbica 3D U-net (usada por el reciente programa de detección LST-AI).

Como evolución de la arquitectura U-net surgió la nnU-Net [14] que, como se ha dicho anteriormente, es la que se emplea para realizar este proyecto. Esta arquitectura es de suma importancia, pues es capaz de adaptarse de forma automática a las características de las imágenes de entrada, autoajustando la arquitectura, los hiperparámetros y el preprocesamiento del modelo a las características del conjunto de datos, eliminando así la elección manual de hiperparámetros, evitando sesgos y fallos humanos. Debido a la adaptabilidad y buenos resultados, además de tener un uso más accesible para los investigadores, la red nnU-net se emplea en muchos estudios de diversas temáticas. Esta arquitectura ha demostrado tener una gran flexibilidad y rendimientos superiores en challenges internacionales como por ejemplo **MSSEG2 challenge**. Los retos internacionales como el que se acaba de citar, u otros como **ISBI challenge** o **MSLesSeg challenge** impulsan la investigación para desarrollar nuevos modelos basados en

aprendizaje profundo (deep learning, DL) que sean capaces de superar las métricas de modelos anteriores.

A continuación se introducen algunos análisis recientes que se han llevado a cabo siguiendo distintos métodos, pero antes de ello se adjunta una tabla resumen con las arquitecturas más empleadas en estos modelos y sus características principales:

Cuadro 2.1: Comparativa de arquitecturas para segmentación de lesiones.

Arquitectura	Tipo de segmentación	Características
U-Net	Segmentación total	Arquitectura clásica precisa
3D U-Net	Segmentación total	Incorpora contexto volumétrico
nnU-Net	Total y longitudinal	Ajuste automático y flexible
LST-AI	Total y longitudinal	Ensamble de múltiples redes

En [23], se define un modelo basado en preentrenamiento autosupervisado y generación de lesiones sintéticas (con el fin de solventar el problema de escasez de datos). El modelo está basado en la red neuronal VNet. Las métricas obtenidas con esta combinación de preentrenamiento autosupervisado y generación de lesiones sintéticas evaluándolo sobre el conjunto MSSEG-2 son DICE promedio de $56.15\% \pm 7.06$ y un F1 Score de $56.69\% \pm 9.12$, superando otros métodos como Coact y SNAC. Este modelo es especialmente bueno para lesiones de bajo contraste y de tamaño pequeño.

En 2024, en el análisis [27] los creadores de LST introducen la herramienta LST-AI. Consiste en un conjunto de tres 3D U-Nets donde el número de capas de supervisión profunda difiere. Una red tiene una sola capa de supervisión profunda, y las otras dos redes tienen dos capas, con el fin de permitir variabilidad en el conjunto de predicciones. Para evaluar su rendimiento, lo compararon con otros cinco métodos de segmentación, que son LST-LGA, LST-LPA, nnU-Net, SAMSEG y DeepLesionBrain. Esta herramienta supera a los otros modelos, excepto en la métrica precisión, que es superado por nnU-Net. Los autores remarcan lo difícil que resulta detectar lesiones pequeñas ($< 10mm^3$).

El modelo [28], realiza una interpolación calibrada entre fragmentos de imagen, mitigando ruido en los bordes de las lesiones y mejorando la coherencia espacial global. Esto es muy útil para análisis longitudinales, donde la comparación entre imágenes requiere alineación precisa y consistencia volumétrica. Obtiene un DICE cercano al 60 %.

El modelo desarrollado en [2] trata sobre la falta de estandarización entre protocolos de adquisición de MRI y el etiquetado manual de las lesiones. Desarrollan modelos que son capaces de corregir estas etiquetas incorrectas, con el fin de minimizar el ruido. Además, entran los modelos en distintos hospitales sin necesidad de compartir los datos directamente, preservando

la privacidad de las personas.

El estudio realizado en [9] incorpora mecanismos de atención, permitiendo hacer focus en regiones específicas del cerebro, aumentando la precisión y reduciendo falsos positivos. Este modelo lo avalúan en dos conjuntos de datos diferentes, y cada uno de ellos obtenidos de máquinas de escáner distintas y en localizaciones distintas, con el fin de evitar sesgos. El modelo se creó comparando dos imágenes, la baseline y la follow-up, como se hace en este proyecto. Las métricas obtenidas también son competitivas.

Asimismo, el estudio [1] combina datos radiómicos tradicionales con técnicas de DL, mejorando la correlación con biomarcadores clínicos. LLegan a la conclusión de que la fusión de los datos obtenidos a través de la radiómica (como Concentration Rate y Rényo Entropy) con CNN mejora la segmentación de imágenes, obteniendo un DICE alrededor del 74 %.

También, en el artículo [8] se comparan pares de imágenes en secuencias FLAIR en dos tiempos distintos, (FLAIR baseline vs FLAIR follow-up). Las imágenes fueron tomadas en colaboración con el Hospital Clínic de Barcelona. En el momento de su publicación obtuvieron métricas que ocupaban el cuarto puesto en el challenge MSSEG2. El procedimiento seguido en este proyecto de memoria es parecido, pero tiene varios puntos bien diferenciados, el dataset de entrenamiento es totalmente distinto al empleado en el artículo citado, y la metodología es distinta debido a las limitaciones computacionales que explicadas más adelante. Además, se introducen técnicas de postprocesado relacionadas con la homología persistente.

3. Nuevas líneas de investigación

Recientemente hay un área de Matemáticas que está aportando avances muy importantes en la segmentación de imágenes. Se trata del TDA (Análisis de Datos Topológicos), en concreto, la rama homología persistente (Persistent Homology, PH). Es muy útil ya que aporta información veraz sobre el número de lesiones detectadas en una imagen. Esto sirve de ayuda para poder descartar aquellas lesiones que son ruido, obteniendo mejores resultados en las métricas de los modelos de segmentación.

En el artículo [13] introducen el algoritmo P-Count, que consiste en, a través de la PH, contar de forma robusta el número de lesiones que se encuentran en la sustancia blanca cerebral. Este algoritmo halla la persistencia de las componentes conexas, filtrando el ruido en función de la persistencia que se haya obtenido. Se centran en la persistencia en el tiempo, de forma que la persistencia de un punto p será la diferencia entre el momento en el que “muere”, y el momento en el que “nace” ese punto.

En el artículo [16] se comparan técnicas relacionadas con teoría de grafos y PH (mediante las curvas de Betti, que son invariantes topológicos), para hallar rasgos diferenciadores en las

imágenes por RM de personas con EM y personas que no la tienen. Se llega a la conclusión de que la técnica que da mejores resultados es la basada en la homología persistente.

Capítulo 3

Métodos y recursos

1. Conjunto de datos ImaginEM

El proyecto parte de la base de datos formada por imágenes de 349 IDs (clave primaria única asociada a una persona) de personas con EM, con un rango de edad de 50 ± 10 años y compuesta aproximadamente por un 60/70% de mujeres. Para cada uno de estos IDs se les ha tomado imágenes por resonancia magnética (RM) cerebrales, en dos momentos temporales distintos, con una diferencia entre 1 y 3 años, esto es lo que se conoce como estudio longitudinal de RM de un mismo sujeto. Estas imágenes se han tomado desde el Hospital Clínic de Barcelona, por parte de los grupos de investigación ImaginEM (Imagen avanzada en enfermedades neuroinmunológicas) e IDIBAPS (Institut d'Investigacions Biomèdiques August Pi i Sunyer). Para cada uno de estos IDs existen dos imágenes en secuencia FLAIR (FLAIR baseline y FLAIR followup) y dos imágenes en secuencia T1 (T1 baseline y T1 followup). Además, para cada ID existe una imagen que contiene las máscaras de segmentación manual (ground truth, GT) realizada por los radiólogos especialistas, donde las lesiones nuevas están etiquetadas con un 2, las lesiones estables con un 1, y las zonas que no son lesiones están etiquetadas con un 0, conocidas como background. A continuación se muestra un ejemplo de esta segmentación manual,

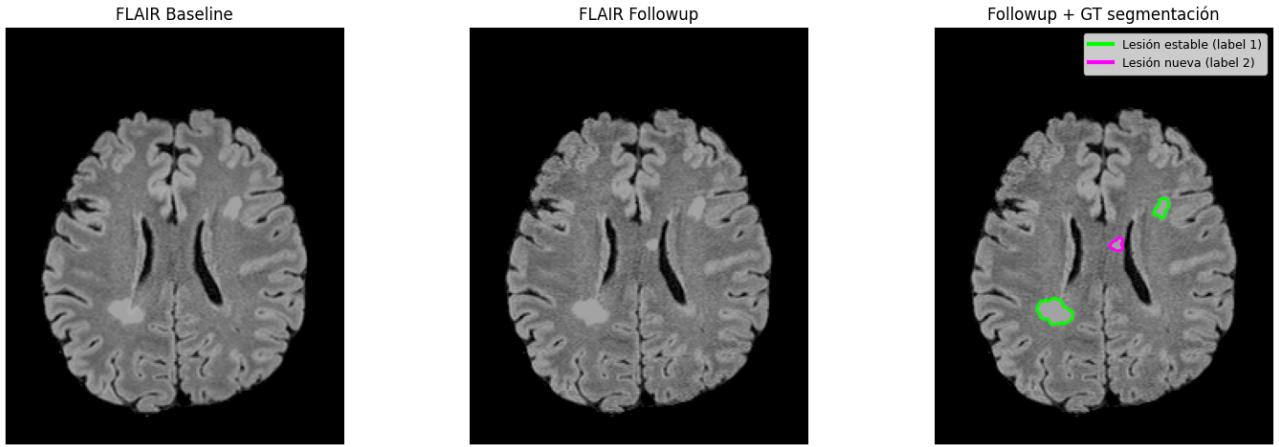


Figura 3.1: Máscara de segmentación manual. Las lesiones rodeadas de verde corresponden con lesiones estables, las rosas con lesiones nuevas.

Las cinco imágenes que hay por cada ID, (FLAIR baseline, FLAIR followup, T1 baseline, T1 followup y la imagen con la segmentación manual), se encuentran en formato NIfTI (.nii.gz) (Neuroimaging Informatics Technology Iniciative). Este formato permite establecer un marco común de recogida y volcado de datos en las imágenes obtenidas por RM, permitiendo de esta forma el uso general y extendido de las imágenes en toda la comunidad científica. A lo largo de esta memoria, se nombrará al conjunto formado por las imágenes para cada uno de los 349 IDs como “dataset ImaginEM”. En este conjunto los datos están anonimizados, impidiendo la revelación de la identidad de cada ID. Además, entre la UOC (Universitat Oberta de Catalunya) y el IDIBAPS existe un comité de ética, que permite el uso de las imágenes siempre que estén anonimizadas.

1.1. Preprocesamiento imágenes RM ImaginEM

Como se comentó anteriormente en la [Sección 4](#), la CNN nnU-Net necesita que las imágenes por RM tengan unos formatos específicos para que puedan ser procesados por esta red. Para ello, debe realizarse un preprocesamiento de forma externa. Las imágenes longitudinales por RM de cada sujeto del dataset ImaginEM empleadas para entrenar los modelos de este proyecto han pasado todas ellas por varios procesos de preprocesamiento, realizados por parte de los equipos de investigación ImaginEM e IDIBAPS. Las técnicas que se realizan son:

- Se define la alineación espacial común del MNI (Montreal Neurological Institute). De esta forma todas las imágenes FLAIR del dataset ImaginEM están alineadas con el sistema de coordenadas MNI, permitiendo la generalización de los modelos a otros datasets.
- Skull-stripping: se aplica el método basado en deep learning, HD-BET (Brain Extraction

Tool) para seleccionar de las imágenes únicamente el tejido cerebral deseado.

- Corrección inhomogeneidades: se aplica el algoritmo N4 para corregir ciertos bias, obteniendo imágenes estándares y consistentes.

Una vez que finaliza este proceso, las imágenes están listas para realizar entrenamientos a través de nnU-Net, incluidos los preprocesamientos integrados de nnU-Net. En los siguientes apartados se desarrollan estudios estadísticos sobre los datasets ImaginEM y MSSEG2.

1.2. Estudio estadístico del conjunto de datos ImaginEM

Realizando un estudio estadístico de este conjunto de datos se obtiene que los 349 IDs han desarrollado todas nuevas lesiones, con una media de 6.32 ± 10.19 lesiones nuevas. Esto hace que el conjunto de datos ImaginEM esté desbalanceado, respecto a IDs con nuevas lesiones vs IDs sin nuevas lesiones. Por ello, el modelo entrenado sobre estos datos puede sufrir sobreajuste, aumentando el número de falsos positivos. Debido a esto, en el proyecto se tiene especial cuidado con las métricas obtenidas y, adicionalmente, se evalúa el modelo entrenado en el dataset externo MSSEG2, que presenta IDs que no desarrollan lesiones nuevas. A continuación se adjunta una tabla donde se muestran las métricas comentadas,

Métrica	Lesiones estables (label=1)	Lesiones nuevas (label=2)
Número de lesiones [n]	$96,87 \pm 71,58$	$6,32 \pm 10,19$
Volumen total [mm^3]	$10693,86 \pm 12887,26$	$906,33 \pm 1709,52$
Casos sin lesiones [n]	0	0

Cuadro 3.1: Resumen estadístico del conjunto de datos ImaginEM.

Se observa que la desviación estándar es mayor que la media, lo que indica que la distribución es fuertemente asimétrica, es decir, hay muchos IDs que presentan pocas lesiones nuevas, pero hay algunos IDs que presentan muchas lesiones nuevas. Poniendo atención en las métricas obtenidas en las lesiones con etiqueta 1 (lesiones estables), se obtiene que la media por cada ID es de 96 ± 71 lesiones, pareciendo indicar que se trata de una cohorte con EM crónica, donde el número de lesiones estables es mayor que el de nuevas. En las siguientes gráficas se muestran estas distribuciones asimétricas.

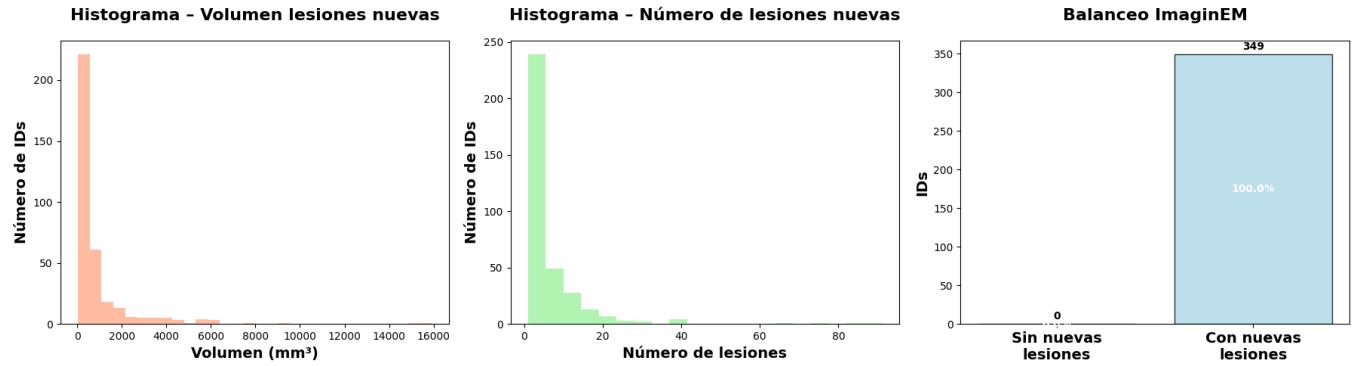


Figura 3.2: Gráficas estudio estadístico del conjunto ImaginEM

Se observa que la mayoría de los IDs tienen alrededor de 5 lesiones nuevas, y la mayoría de sus volúmenes no superan los 1000 mm^3 .

2. Conjunto de datos MSSEG2

El conjunto MSSEG2 contiene por cada ID los siguientes elementos: dos imágenes en secuencias FLAIR, (FLAIR baseline y FLAIR followup con una diferencia temporal entre la primera y la segunda de uno a tres años), y cinco máscaras de segmentación manual, en las que se etiqueta con 0 las zonas sin lesiones, y con 1 las zonas con nuevas lesiones. Cuatro de las cinco máscaras se corresponden a la segmentación manual realizada por cada uno de los cuatro expertos radiólogos que hicieron las segmentaciones. La quinta segmentación manual consiste en la máscara acordada finalmente por los cuatro radiólogos. Esta máscara consensuada es la que se ha usado en este proyecto para evaluar el modelo en MSSEG2. Los creadores de este conjunto ponen a disposición del público el dataset que usaron de entrenamiento, formado por 40 IDs, éste es el que se ha empleado para evaluar el modelo del proyecto. Como en el conjunto ImaginEM, las imágenes están en formato NIfTI (.nii.gz), y la información está anonimizada.

2.1. Estudio estadístico del conjunto de datos MSSEG2

Tras realizar un análisis estadístico sobre este conjunto, la primera diferencia con ImaginEM es que no todos los IDs desarrollan nuevas lesiones. En la siguiente tabla se pueden observar las métricas obtenidas,

Métrica	Lesiones nuevas (label=1)
Número de lesiones nuevas [n]	$3,90 \pm 6,27$
Volumen total nuevas [mm^3]	$558,88 \pm 1395,47$
Casos sin lesiones nuevas [n]	11

Cuadro 3.2: Resumen estadístico del conjunto de datos MSSEG2.

Las métricas indican que la media de lesiones nuevas es de 3.90 ± 6.27 , y el volumen medio de las lesiones nuevas de $558.88 \pm 1395.47 \text{ mm}^3$. En ambos casos la desviación estándar supera casi el doble de la media, lo que indica que la distribución de las lesiones es fuertemente asimétrica. Esto se puede ver en los diagramas siguientes, en los que se observa un sesgo positivo en el volumen de las lesiones nuevas y en el número de las lesiones. En el último diagrama se muestra el desbalanceamiento que existe entre los IDs que no desarrollan nuevas lesiones y los que sí (27.5 % vs 72.5 %).

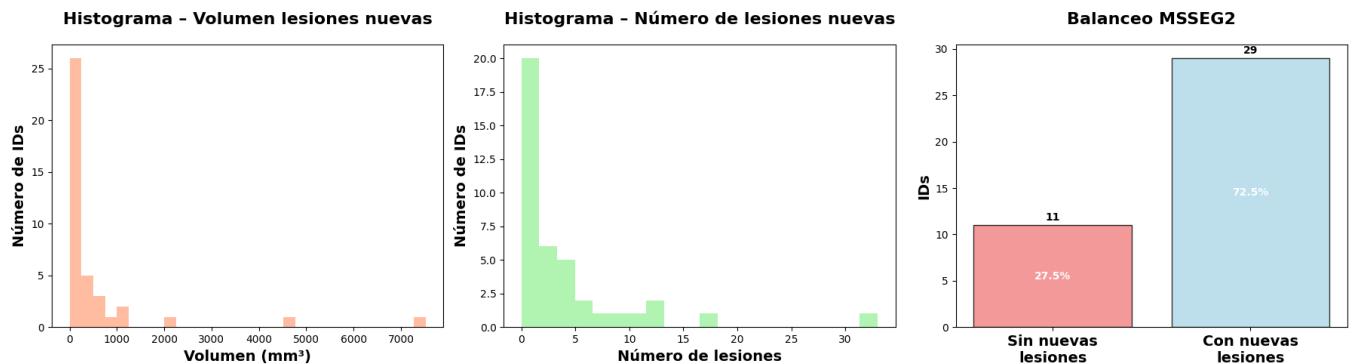


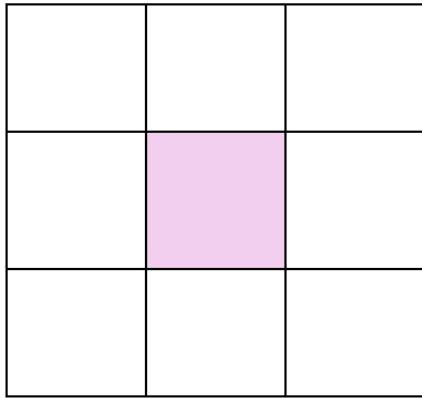
Figura 3.3: Gráficas estudio estadístico del conjunto MSSEG2

En el primer diagrama se observa que la mayoría de IDs presentan lesiones nuevas con volúmenes menores a 500mm^3 , y en el segundo diagrama se muestra que la mayoría de IDs no tiene más de cinco lesiones nuevas.

3. Definiciones previas

Definición Como se explica en [18, p. 707], un **voxel** es una representación en 3D, equivalente al píxel en 2D. Al tener tres dimensiones estamos tratando con un elemento volumétrico. Cabe destacar que en este proyecto en un mismo voxel únicamente puede darse una etiqueta en GT, puesto que las etiquetas, tanto en ImaginEM, como en MSSEG2, no son soft labels (lista de probabilidades para cada una de las clases por cada voxel). A continuación se muestra la diferencia entre representar un píxel en 2D, y hacerlo en 3D (voxel).

Representación de los píxeles en 2D



Representación de los píxeles en 3D

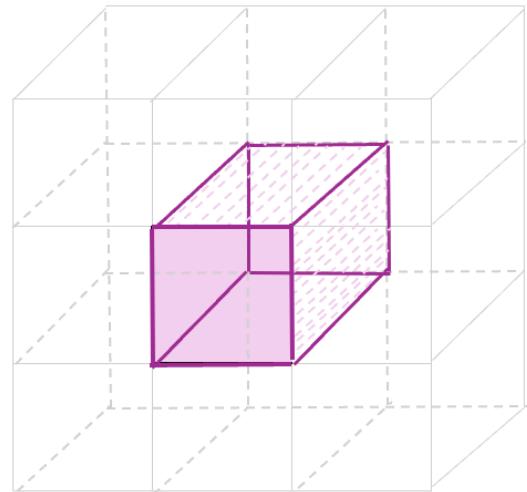


Figura 3.4: Representación píxel vs representación voxel

Un aspecto clave es que **los voxels de cada imagen no tienen dimensiones estandarizadas**. Su espaciado en las direcciones x , y , z , varía según el escáner, protocolo y configuración, haciendo que los volúmenes de los voxels de imágenes tomadas con escáneres/protocolos/configuraciones distintas difieran entre sí.

Estudiando las características de los voxels de las imágenes de ImaginEM se obtiene que únicamente presenta tres combinaciones distintas de coordenadas x,y,z , que dan lugar a volúmenes similares, lo que demuestra que todas las imágenes se tomaron con un protocolo unificado en el Hospital Clínic de Barcelona. En la siguiente tabla se muestran las combinaciones,

ImaginEM	dim_x (mm)	dim_y (mm)	dim_z (mm)	voxel_vol (mm ³)	shape_x (nº voxel)	shape_y (nº voxel)	shape_z (nº voxel)
Combinación 1	0.86	0.859375	0.859375	0.63513184	208	256	256
Combinación 2	0.86	0.9375	0.9375	0.75585943	208	256	256
Combinación 3	0.94	0.9375	0.9375	0.82617190	240	256	256

Cuadro 3.3: Espaciado voxel y dimensiones en ImaginEM.

En cambio, en los 40 IDs de MSSEG2 se encuentran hasta 33 combinaciones de estas coordenadas, con mucha variación entre las dimensiones de un voxel y otro. Esto es debido a que en MSSEG2 se han tomado las imágenes desde centros distintos y con escáneres distintos, lo que hace que haya una gran variabilidad, y sea un dataset idóneo para comprobar la capacidad de generalización del modelo entrenado.

Definición Una lesión estará contenida en uno o más voxels conexos entre sí, que hayan sido etiquetados con la misma clase (lesión estable etiqueta 1, lesión nueva etiqueta 2).

Representando en 3D en forma de nube de puntos el centro de cada voxel de cada tipo de lesión (a partir de la máscara manual), y con una muestra reducida de puntos grises la zona sin lesiones (para no saturar la visualización), las lesiones se ven de esta forma,

Lesiones estables y nuevas en 3D (ImaginEM)

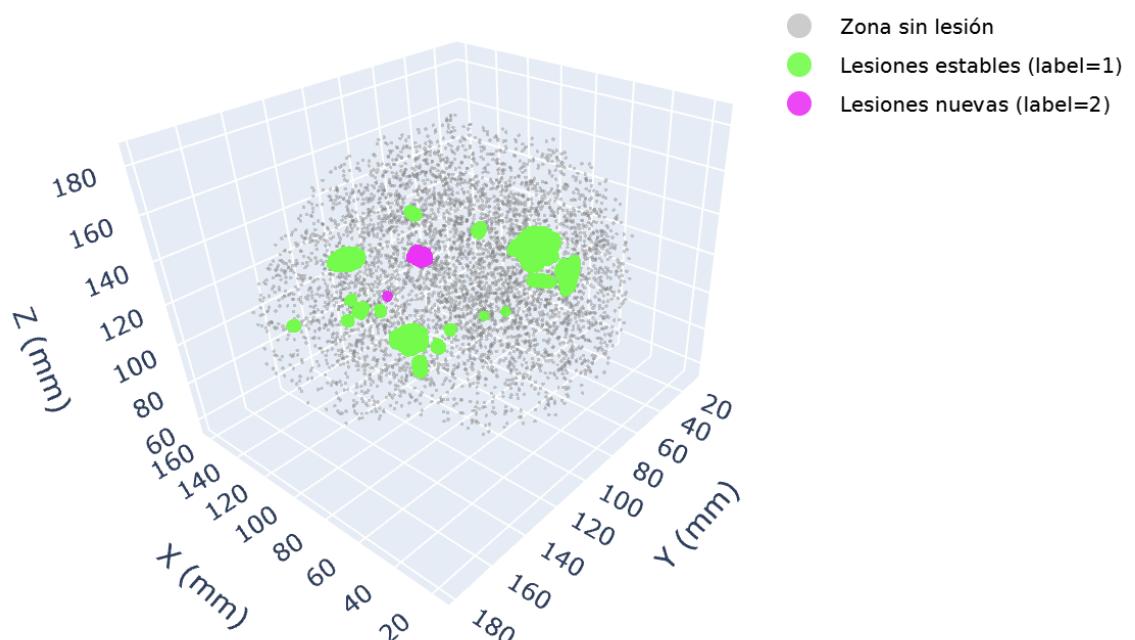


Figura 3.5: Representación 3D de lesiones en voxels

En la siguiente imagen se muestra cómo se distinguen los voxels haciendo zoom,

Lesiones estables y nuevas en 3D (ImaginEM)

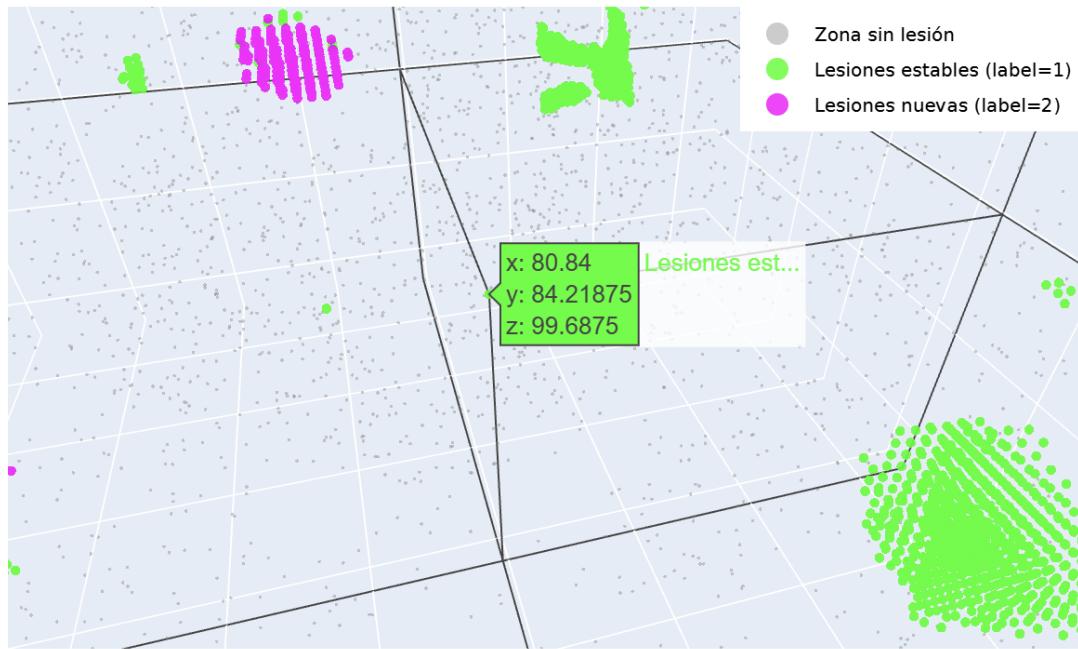


Figura 3.6: Voxelés de lesiones

Definición Una métrica a nivel **voxel-wise** se obtiene comparando la etiqueta predicha de cada voxel con la etiqueta que se le ha asociado en GT. Así, se obtienen conteos voxel a voxel, identificando cada uno de ellos como TP, FP, FN y TN, y hallando a partir de esto las métricas Dice, F1, sensibilidad, recall...

Definición Una métrica a nivel **lesion-wise** evalúa la capacidad de detección de las lesiones, tomando por lesión un conjunto de voxels conectados entre sí que tengan la misma etiqueta en GT y en la predicción. Por ejemplo, una lesión será un TP si para al menos un voxel coincide la etiqueta de la predicción con la de la segmentación manual. El resto de voxels de la lesión se considerarán FP o FN. Una vez que se obtienen los TP, FP y FN para cada lesión, se obtienen métricas como la sensibilidad y la precisión.

Definición Una métrica a nivel **ID-wise** estudia el rendimiento del modelo por cada ID del conjunto de datos. Agrega la información de todas las lesiones del GT, todas las lesiones que ha detectado el modelo y las que no, y a partir de este cómputo total se obtienen las distintas métricas.

3.1. Métricas

En los problemas de clasificación, como el de este proyecto (lesión estable vs lesión nueva; zona sin lesión vs zona con lesión), para medir el rendimiento de los distintos modelos entrenados se comparan las predicciones generadas frente al GT.

En los problemas de clasificación binarios, con dos clases: P, la clase positiva, y N, la clase negativa, las métricas más usadas son:

- Verdadero positivo (TP): es el número de clasificaciones correctas de la clase positiva P.
- Verdadero negativo (TN): es el número de clasificaciones correctas en la clase negativa N.
- Falso negativo (FN): es el número de clasificaciones incorrectas de la clase positiva, clasificada como negativa.
- Falso positivo (FP): es el número de clasificaciones incorrectas de la clase negativa, clasificada como positiva.
- Precisión: $\frac{TP}{TP+FP}$
- Tasa de verdaderos positivos (TPR): $\frac{TP}{TP+FN}$
- Tasa de falsos positivos (FPR): $\frac{FP}{FP+TN}$
- Recall=sensibilidad=TPR
- F1: $F1 = \frac{\text{Precisión}\cdot\text{Recall}}{\text{Precisión}+\text{Recall}}$
- Dice: $Dice = \frac{2TP}{2TP+FP+FN}$

En el proyecto se han analizado los distintos modelos a nivel de voxel, de lesión y de ID.

Las métricas usadas en las evaluaciones de los modelos realizadas en el conjunto de datos MSSEG2 son: a nivel de voxel el número de FP y el Dice, y a nivel de lesión el número de FP y F1.

4. Procedimiento

Como se ha ido comentando a lo largo de esta memoria, el proyecto se basa en la CNN nnU-Net v2. En [14] se presentó nnU-Net. Realiza de forma automática el preprocesamiento de las imágenes del conjunto de datos a estudiar, la configuración de la arquitectura de la

red neuronal de entrenamiento y la elección autónoma de parámetros de entrenamiento como el learning rate (lr), el número de épocas, el optimizador a usar... Esto hace que el manejo de la red sea más accesible para los usuarios, ya que no se necesita un experto tan elevado, al no tener la necesidad de decidir qué parámetros de entrenamiento elegir o qué técnicas de preprocesamiento aplicar (normalización, data augmentation, oversampling...), evitando la toma de decisiones equivocadas a la hora de elegir estos parámetros. En este artículo evaluaron el rendimiento de la red nnU-Net en 11 retos internacionales de segmentación de imágenes médicas, que incluían 23 conjuntos de datos y 53 tipos de segmentación, y se obtuvo que la red nnU-Net conseguía en la mayoría de casos mejores resultados que otros procesos basados en diferentes redes neuronales, posicionándose como el framework más flexible y competitivo en segmentación médica. Por defecto, los entrenamientos en esta red se ejecutan durante 1000 épocas con un lr inicial de 0,01, pero configurado de forma que va disminuyendo en cada época (poly lr decay). La red nnU-Net v2 es una mejora que realizaron los creadores de nnU-Net en 2023, [GitHub creador nnU-Net](#), por ello es la que se emplea en el proyecto.

El dataset externo MSSEG2 únicamente tiene por cada ID las imágenes FLAIR baseline y FLAIR followup, es por ello por lo que en este proyecto para entrenar los modelos únicamente se toman los canales FLAIR baseline y FLAIR followup del conjunto de datos ImaginEM.

4.1. Estrategias entrenamiento

La idea inicial del proyecto fue entrenar un modelo con el dataset ImaginEM, sin realizar ninguna modificación en la configuración de la nnU-Net v2. Esto es, iba a realizarse un entrenamiento con validación cruzada en cinco folds y configuración fullres 3D en cada fold durante 1000 épocas. Las limitaciones de computación producidas por la ausencia de servidor/máquina virtual/tarjeta gráfica dedicada con RAM superior a 12GB, hizo inviable este entrenamiento. La única opción viable económicamente tras la ausencia de estos recursos, fue realizar el entrenamiento con una suscripción a Google Colab Pro+, que permitía el uso de tarjetas gráficas A100, pero con una RAM compartida entre todos los usuarios (lo que en realidad correspondía con una GPU de 16/17.5GB). Cada entrenamiento de 1000 épocas tardaba 40 horas en ejecutarse para un solo fold, con la limitación de que las sesiones de Google Colab Pro+ duran un máximo de 24 horas. Esto significaba que para hacer el entrenamiento con validación cruzada en 5 folds, se necesitaban 200 horas de computación para un solo modelo, lo que equivale a más de 8 días seguidos. Era totalmente inviable porque el tiempo requerido de computación no permitía hacer ninguna modificación o mejora en base a los resultados que se obtuvieran en las evaluaciones del modelo.

Por ello, se descartó la técnica de la validación cruzada en el entrenamiento del modelo, y se decidió realizar una evaluación del modelo sobre el propio dataset ImaginEM, lo que se

conoce como holdout. Se tomó un 70% del dataset para entrenar, un 15% para validar el entrenamiento y el 15% restante para evaluar el modelo, a esto se le conoce como test-split. Al tener el dataset ImaginEM una muestra de IDs grande, 349, este procedimiento es robusto. Cabe destacar que esta división de ImaginEM en tres subconjuntos (train, validation y test) se hizo de forma que fueran disjuntos, es decir, que ninguno de los conjuntos tuviera elementos en común ($\text{train} \cap \text{validation} = \emptyset$, $\text{train} \cap \text{test} = \emptyset$ y $\text{validation} \cap \text{test} = \emptyset$), con el fin de garantizar la independencia de cada muestra y la validez y robustez de las métricas obtenidas. Además, para evitar sesgos introducidos al realizar esta partición, se hizo de forma aleatoria. El reparto quedó de la siguiente forma: train contenía información de 244 IDs, validation contenía información de 52 IDs y test contenía información de 53 IDs, sin tener ninguno de ellos IDs en común.

Una vez descartada la técnica de cross validation, seguía existiendo el problema de que el entrenamiento se iba a realizar durante 1000 épocas (40 horas de computación), por ello, el siguiente paso fue configurar entrenamientos personalizados en los que se podía fijar el número de épocas en la nnU-Net v2. Esta información se obtuvo del repositorio GitHub: [nnUNetTrainer-Xepochs](#).

Para saber qué número de épocas se fijaba en el modelo final, se configuraron 4 modelos distintos:

- Modelo con 20 épocas.
- Modelo con 50 épocas.
- Modelo con 100 épocas.
- Modelo con 250 épocas.

Con el fin de, una vez hubieran sido evaluados en test-split y MSSEG2, comparar sus métricas (comentadas en el [Capítulo 4](#)) para elegir el modelo que mejor rendimiento ofreciera.

A continuación se muestran las gráficas de progreso de cada uno de los entrenamientos de los 4 modelos distintos:

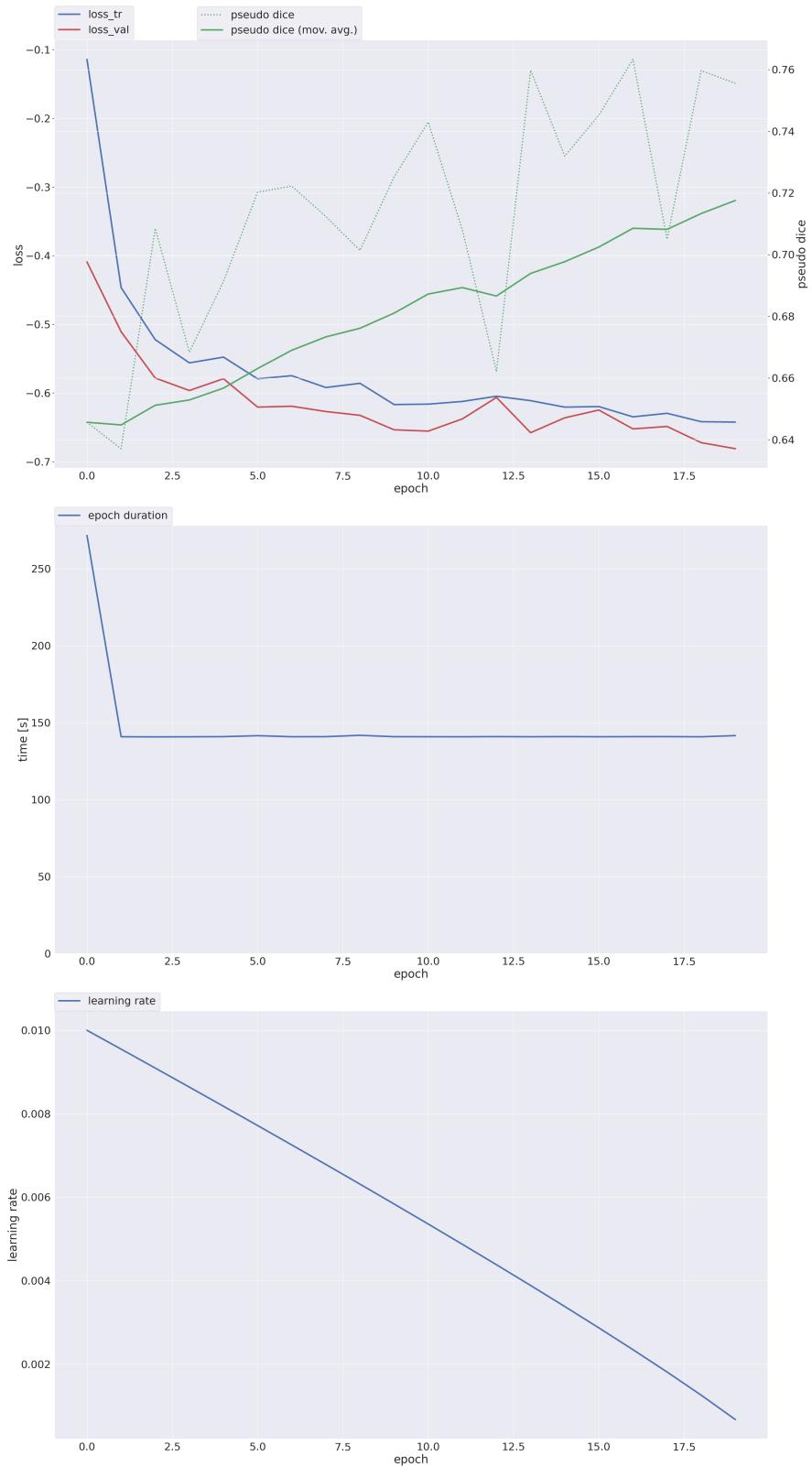


Figura 3.7: Entrenamiento modelo HOLDOUT con 20 épocas.

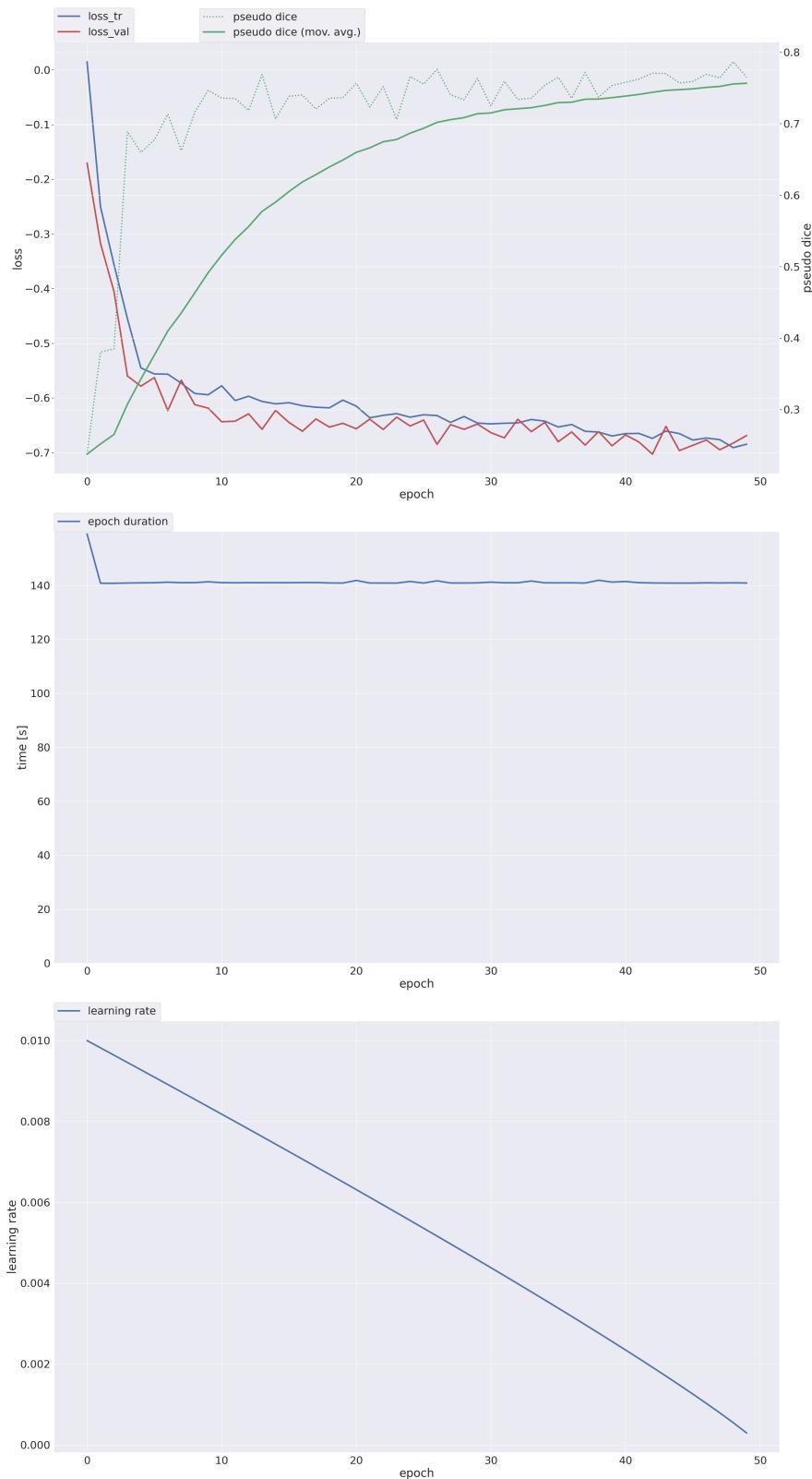


Figura 3.8: Entrenamiento modelo HOLDOUT con 50 épocas.

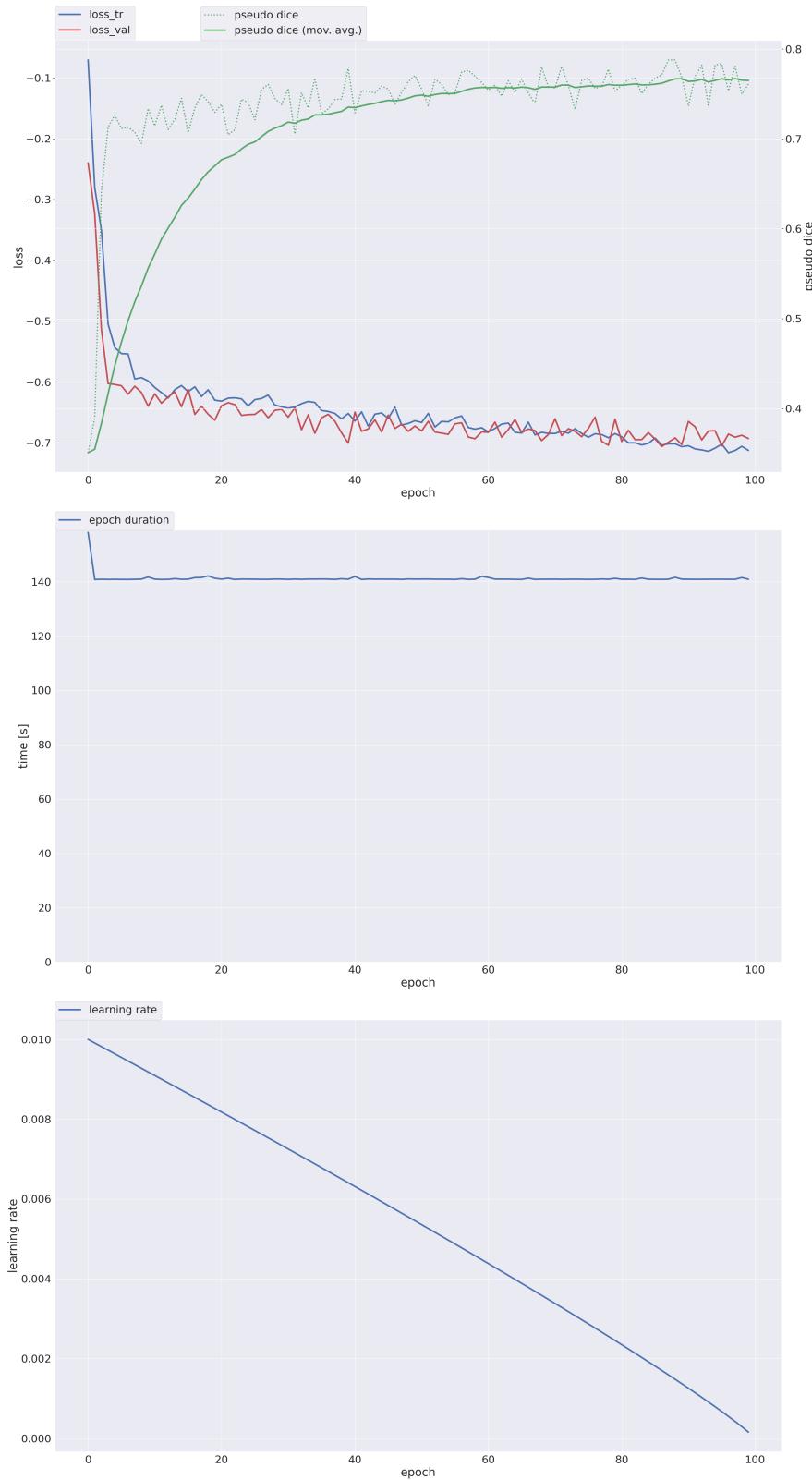


Figura 3.9: Entrenamiento modelo HOLDOUT con 100 épocas.

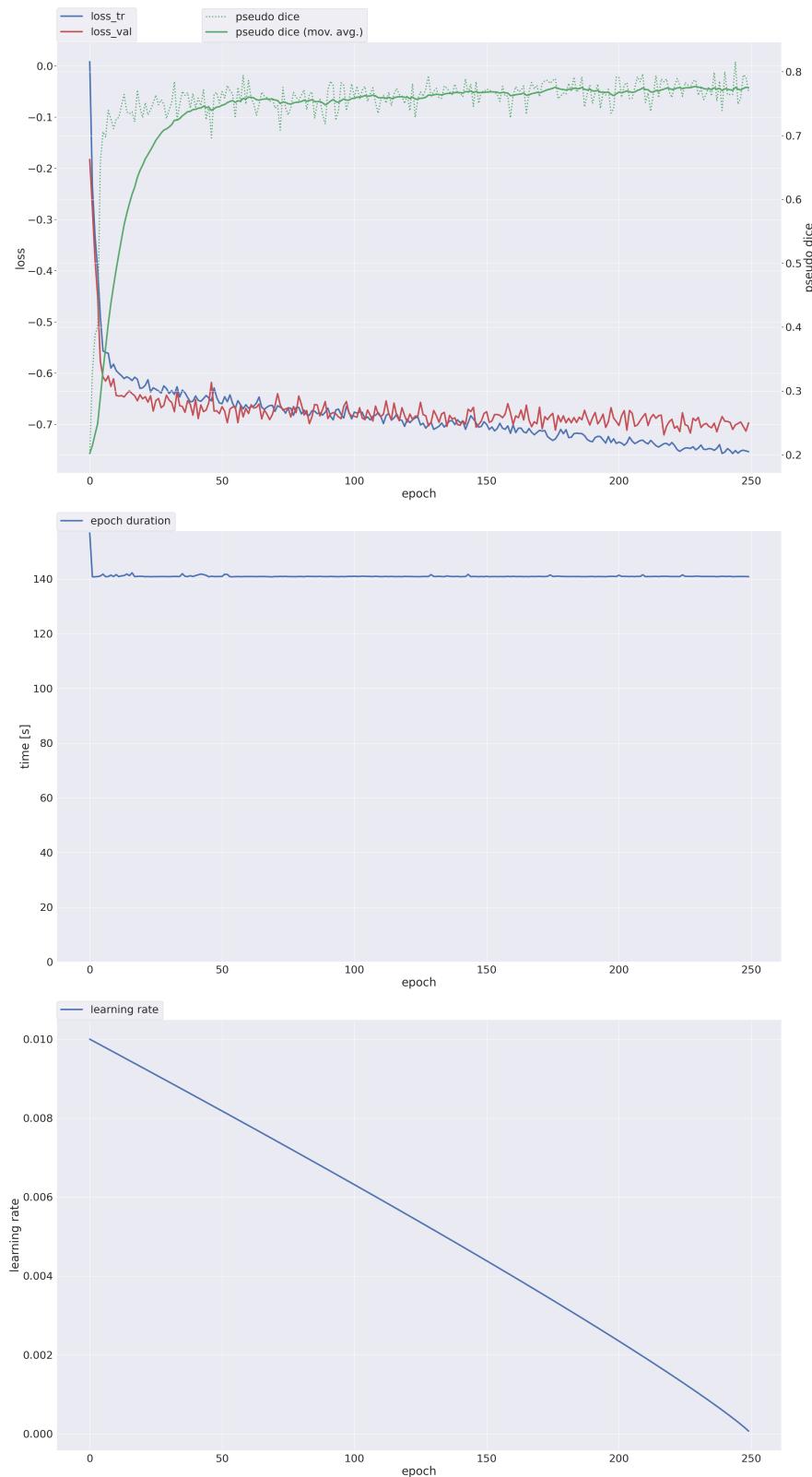


Figura 3.10: Entrenamiento modelo HOLDOUT con 250 épocas.

Observando las distintas curvas de entrenamiento se puede concluir que el entrenamiento más equilibrado es el de 100 épocas, ya que la curva de pseudo-Dice de validación se estabiliza en un valor alto sin que la pérdida de validación empiece a degradarse y, al mismo tiempo, es donde el número de falsos positivos en test split y en MSSEG2 es mínimo o cercano al mínimo (como se verá en el [Capítulo 4](#)). Prestando atención en las curvas del entrenamiento de 20 épocas, éstas aún son inestables al final. Para el de 50 épocas, la loss de train/val sigue bajando y el pseudo-Dice sube respecto al de 20 épocas, pero todavía hay oscilaciones visibles. En el de 100 épocas las curvas de loss se aplanan y el pseudo-Dice de validación se mantiene alto y mucho más estable; el lr ya es bajo, permitiendo refinamiento sin grandes saltos. Es el que mejor comportamiento tiene. En el entrenamiento con 250 épocas la loss/pseudo-Dice de validación prácticamente no mejora.

En conclusión, a partir de las 100 épocas las curvas apenas cambian, por lo que el modelo con 100 épocas ya ha capturado la estructura general del dataset, y son suficientes épocas para no alcanzar un sobreentrenamiento. Por ello, el modelo final que se escoge es el de 100 épocas.

Capítulo 4

Resultados

En este capítulo se comparan las métricas obtenidas en cada uno de los 4 modelos (20 épocas, 50 épocas, 100 épocas y 250 épocas), evaluados en test-split (53 casos) y en el dataset externo MSSEG2 (40 casos).

La evaluación en test-split contempla las clases 0 (sin lesión), 1 (lesión estable) y 2 (lesión nueva). Las métricas “Dice new”, “Prec. new” (precision), “Rec. new” (recall), “F1 new” corresponden a la predicción para las clases con etiqueta 2. Las métricas “Prec. stable” (precision), “Rec. stable” (recall), “F1 stable” corresponden a la predicción para las clases con etiqueta 1.

La evaluación en MSSEG2 contempla las clases sin lesión nueva (etiquetas 0 y 1 en ImaginEM, etiqueta 0 en MSSEG2) vs las que tienen lesión nueva (etiqueta 2 en ImaginEM, etiqueta 1 en MSSEG2). Como MSSEG2 no tiene etiqueta para lesiones estables, la etiqueta 1 de ImaginEM se tiene que considerar dentro de zonas sin lesión nueva (es decir, se colapsan las etiquetas 0 y 1 de ImaginEM como zona sin lesión nueva). Las métricas “Dice new”, “Prec. new”, “Rec. new”, “F1 new” corresponden a la predicción para las clases con lesión nueva. Las métricas “Prec. no new”, “Rec. no new”, “F1 no new” corresponden a la predicción para las clases sin lesión nueva.

Como se explicó en [Sección 3](#), las métricas halladas a nivel lesion-wise e ID-wise, tanto para la evaluación en test-split como en la evaluación externa en MSSEG2, se han hallado para las predicciones de lesiones nuevas (etiqueta 2 en ImaginEM y etiqueta 1 en MSSEG2).

1. Evaluación test-split

1.1. Voxel-wise ImaginEM

Epochs	TP	FP	FN	TN	Dice new	Prec. new	Rec. new	F1 new
20	56535	9886	7349	501201	0.8677	0.8512	0.5801	0.6900
50	59235	10778	7300	506408	0.8676	0.8461	0.6078	0.7074
100	54105	8213	9237	518167	0.8611	0.8682	0.5552	0.6773
250	54749	8020	9083	519213	0.8649	0.8722	0.5618	0.6834

Cuadro 4.1: Resultados voxel-wise para la lesión nueva en test-split ImaginEM.

Epochs	Prec. stable	Rec. stable	F1 stable	F1 macro	F1 weighted
20	0.9855	0.7954	0.8803	0.7851	0.8548
50	0.9858	0.8037	0.8855	0.7964	0.8616
100	0.9825	0.8223	0.8953	0.7863	0.8661
250	0.9828	0.8240	0.8964	0.7899	0.8679

Cuadro 4.2: Métricas voxel-wise para la lesión estable y promedios en test-split ImaginEM.

1.2. Lesion-wise ImaginEM

Epochs	GT lesiones	GT detectadas	FP lesiones	Sensibilidad	Precisión
20	389	287	111	0.7378	0.7211
50	389	296	139	0.7609	0.6805
100	389	247	79	0.6350	0.7577
250	389	249	73	0.6401	0.7733

Cuadro 4.3: Resultados lesion-wise en test-split ImaginEM.

1.3. ID-wise ImaginEM

Epochs	Casos	Casos con lesión nueva	Casos detectados	Sensibilidad
20	53	53	52	0.9811
50	53	53	52	0.9811
100	53	53	52	0.9811
250	53	53	53	1.0000

Cuadro 4.4: Resultados ID-wise en test-split ImaginEM.

2. Evaluación MSSEG2

2.1. Voxel-wise MSSEG2

Epochs	TP	FP	FN	TN	Dice new	Prec. new	Rec. new	F1 new
20	33427	691989	26490	1321699294	0.0851	0.0461	0.5579	0.0851
50	37863	280792	22054	1322110491	0.2000	0.1188	0.6319	0.2000
100	34795	86505	25122	1322304778	0.3840	0.2869	0.5807	0.3840
250	28624	121196	31293	1322270087	0.2730	0.1911	0.4777	0.2730

Cuadro 4.5: Resultados voxel-wise para la lesión nueva en MSSEG2.

Epochs	Prec. no new	Rec. no new	F1 no new	F1 macro	F1 weighted
20	0.99998	0.99948	0.99973	0.5424	0.9997
50	0.99998	0.99979	0.99989	0.6000	0.9998
100	0.99998	0.99994	0.99996	0.6920	0.9999
250	0.99998	0.99991	0.99994	0.6364	0.9999

Cuadro 4.6: Métricas voxel-wise para voxels sin lesión nueva y promedios en MSSEG2.

2.2. Lesion-wise MSSEG2

Epochs	GT lesiones	GT detectadas	FP lesiones	Sensibilidad	Precisión
20	154	84	779	0.5455	0.0973
50	154	105	1414	0.6818	0.0691
100	154	78	136	0.5065	0.3645
250	154	70	129	0.4545	0.3518

Cuadro 4.7: Resultados lesion-wise en MSSEG2.

2.3. ID-wise MSSEG2

Epochs	Casos	Casos con lesión nueva	Casos detectados	Sensibilidad
20	40	29	29	1.0000
50	40	29	29	1.0000
100	40	29	28	0.9655
250	40	29	28	0.9655

Cuadro 4.8: Resultados ID-wise en MSSEG2.

Comparando las tablas se observa que en ambos datasets el modelo de 100 épocas es el que mejor equilibrio aporta. Además, es el que mejor Dice posee en MSSEG2.

Como se verá a continuación, en el siguiente paso se aplican varios postprocesados al modelo de 100 épocas, utilizando técnicas de homología persistente (persistent homology, PH) y filtros por volumen (mm^3) (usando los volúmenes de las lesiones segmentadas manualmente del dataset ImaginEM), para analizar qué estrategia es la que reduce más falsos positivos, obteniendo mejores métricas, sin comprometer la sensibilidad.

3. Postprocesado

Una vez que se concluye que el modelo que mejor equilibrio proporciona entre recursos de computación y métricas obtenidas es el de 100 épocas, el siguiente paso es la realización del postprocesado del modelo. Para ello se tomaron cuatro estrategias distintas, basadas en:

1. Volumen.
2. PH.
3. PH+volumen.

4. Volumen+PH.

3.1. Estrategias postprocesado

A continuación se detallan las cuatro estrategias seguidas para realizar el postprocesado.

3.1.1. Postprocesado por volumen de lesiones

Este fue el primer postprocesado llevado a cabo en el proyecto. Primero se halló el volumen en mm^3 de cada lesión nueva de la máscara de segmentación manual del dataset ImaginEM. Una vez que se recopilaron todos los volúmenes, se calcularon sus percentiles 1, 3, 5 y 10. Obteniendo:

- Percentil 1: volumen de $38.0825mm^3$
- Percentil 3: volumen de $45.3738mm^3$
- Percentil 5: volumen de $54.5649mm^3$
- Percentil 10: volumen de $65.1023mm^3$

El siguiente paso fue hallar el volumen para todos los voxels que el modelo de 100 épocas (denominado de ahora en adelante como “100 raw”) había predicho como lesión nueva, y aplicar como filtro el percentil que se había hallado en cada uno de estos casos (percentil 1, percentil 3, percentil 5 y percentil 10). Por ejemplo, si se filtra por el percentil 1, todos aquellos voxels que han sido predichos como lesiones nuevas (etiqueta=2) y tienen un volumen menor al del percentil 1, $38.0825mm^3$, pasan a etiquetarse como 0. Esto se aplica por cada percentil y después se evalúa en los datasets ImaginEM y MSSEG2.

Las métricas que se obtienen en el **test-split de ImaginEM** son:

3.1.1.1. Voxel-wise ImaginEM

Modelo	TP	FP	FN	TN	Dice new	Prec. new	Rec. new	F1 new
100_raw	54105	8213	9237	518167	0.8611	0.8682	0.5552	0.6773
100_P1	52402	7097	9237	518167	0.8652	0.8807	0.5377	0.6677
100_P3	51671	6927	9237	518167	0.8647	0.8818	0.5302	0.6622
100_P5	50685	6647	9237	518167	0.8645	0.8841	0.5201	0.6549
100_P10	49215	6237	9237	518167	0.8641	0.8875	0.5050	0.6437

Cuadro 4.9: Resultados voxel-wise para la lesión nueva en ImaginEM con postprocesado volumen.

Modelo	Prec. stable	Rec. stable	F1 stable	F1 macro	F1 weighted
100_raw	0.9825	0.8223	0.8953	0.7863	0.8661
100_P1	0.9825	0.8223	0.8953	0.7815	0.8648
100_P3	0.9825	0.8223	0.8953	0.7788	0.8641
100_P5	0.9825	0.8223	0.8953	0.7751	0.8631
100_P10	0.9825	0.8223	0.8953	0.7695	0.8616

Cuadro 4.10: Resultados voxel-wise para la lesión estable y promedios en ImaginEM con postprocesado volumen.

3.1.1.2. Lesion-wise ImaginEM

Modelo	GT lesiones	GT detectadas	FP lesiones	Sensibilidad	Precisión
100_raw	389	247	79	0.6350	0.7577
100_P1	389	183	21	0.4704	0.8971
100_P3	389	169	17	0.4344	0.9086
100_P5	389	154	13	0.3959	0.9222
100_P10	389	134	10	0.3445	0.9306

Cuadro 4.11: Resultados lesion-wise en ImaginEM para el modelo de 100 épocas con postprocesado volumen.

3.1.1.3. ID-wise ImaginEM

Modelo	Casos	Casos con lesión nueva	Casos detectados	Sensibilidad
100_raw	53	53	52	0.9811
100_P1	53	53	52	0.9811
100_P3	53	53	52	0.9811
100_P5	53	53	48	0.9057
100_P10	53	53	45	0.8491

Cuadro 4.12: Resultados ID-wise en ImaginEM para el modelo de 100 épocas con postprocesado volumen.

Las métricas que se obtienen en **MSSEG2** son:

3.1.1.4. Voxel-wise MSSEG2

Modelo	TP	FP	FN	TN	Dice new	Prec. new	Rec. new	F1 new
100_raw	34795	86505	25122	1322304778	0.3840	0.2869	0.5807	0.3840
100_P1	33944	83266	25973	1322308017	0.3833	0.2896	0.5665	0.3833
100_P3	33305	81744	26612	1322309539	0.3807	0.2895	0.5559	0.3807
100_P5	32777	81318	27140	1322309965	0.3767	0.2873	0.5470	0.3767
100_P10	32267	80701	27650	1322310582	0.3733	0.2856	0.5385	0.3733

Cuadro 4.13: Resultados voxel-wise para la lesión nueva en MSSEG2 modelo de 100 épocas con postprocesado volumen.

Modelo	Prec. no_new	Rec. no_new	F1 no_new	F1 macro	F1 weighted
100_raw	0.99998	0.99994	0.99996	0.6920	0.99993
100_P1	0.99998	0.99994	0.99996	0.6916	0.99993
100_P3	0.99998	0.99994	0.99996	0.6903	0.99993
100_P5	0.99998	0.99994	0.99996	0.6883	0.99993
100_P10	0.99998	0.99994	0.99996	0.6866	0.99993

Cuadro 4.14: Resultados voxel-wise para voxoles sin lesión nueva y promedios en MSSEG2 modelo de 100 épocas con postprocesado volumen.

3.1.1.5. Lesion-wise MSSEG2

Modelo	GT lesiones	GT detectadas	FP lesiones	Sensibilidad	Precisión
100_raw	154	78	136	0.5065	0.3645
100_P1	154	56	43	0.3636	0.5657
100_P3	154	49	32	0.3182	0.6049
100_P5	154	44	30	0.2857	0.5946
100_P10	154	40	28	0.2597	0.5882

Cuadro 4.15: Resultados lesion-wise en MSSEG2 para el modelo de 100 épocas con postprocesado volumen.

3.1.1.6. ID-wise MSSEG2

Modelo	Casos	Casos con lesión nueva	Casos detectados	Sensibilidad
100_raw	40	29	28	0.9655
100_P1	40	29	26	0.8966
100_P3	40	29	24	0.8276
100_P5	40	29	23	0.7931
100_P10	40	29	22	0.7586

Cuadro 4.16: Resultados ID-wise en MSSEG2 para el modelo de 100 épocas con postprocesado volumen.

Observando las métricas obtenidas, el percentil que mejores resultados aporta, con un equilibrio entre falsos positivos, precisión y sensibilidad es el percentil 3, es decir, tomar un volumen mínimo de 45.3738mm^3 . El percentil 1 aún tiene muchos falsos positivos y, si se toman los percentiles 5 o 10, la sensibilidad se ve fuertemente castigada, ya que disminuye considerablemente en ambos datasets.

3.1.2. Postprocesado por homología persistente

A continuación se va a dar una explicación breve sobre la PH. Es una rama de estudio del Análisis de Datos Topológicos (TDA), dentro de la Topología Algebraica, que se centra en hallar el grupo de homología de los elementos del conjunto de datos. En este proyecto el análisis se realiza sobre H_0 , que corresponde con el grupo de homología de dimensión 0, es decir, las componentes conexas del conjunto de datos. En el caso particular del proyecto son las componentes conexas correspondientes a lesiones cerebrales. Para entender fácilmente el concepto de conexión, un conjunto es conexo si se puede ir de un punto suyo a otro cualquiera trazando una línea sin que esta se salga del contorno del elemento, por ejemplo, la Tierra es conexa. Esta noción se conoce como conexión por caminos, pero no todo elemento que sea conexo por caminos es conexo globalmente, como por ejemplo el Peine del Topólogo, que es conexo por caminos pero no es globalmente conexo. En el caso que atañe en este proyecto vale con la noción de poder ir de un sitio a otro cualquiera del elemento trazando una línea sin levantar el lápiz del papel. En este proyecto se dirá que dos voxels cualesquiera de la imagen por RM son conexos entre sí si alguno de sus vértices, aristas o caras se tocan; en ese caso los voxels conexos pasan a formar parte de la misma lesión, o de la misma zona sin lesión. A esta concepción de conexión se la conoce como 26 conectividad (cada voxel=cubo tiene 6 caras, 8 vértices y 12 aristas, lo que hace un total de 26 elementos posibles por donde puede haber conexión con otro voxel). Se verá a continuación que se han calculado dos invariantes, $n \ bars h0$, que hace una especie de conteo de las manchas brillantes separadas que tiene una lesión en la imagen FLAIR followup, y $sum \ life \ h0$, que observa cuánto duran esas manchas brillantes antes de difuminarse. Las lesiones que son TP tendrán muchas manchas brillantes durante un periodo prolongado de tiempo. Mientras que los puntos que son ruido tendrán pocas manchas y durarán cortos periodos de tiempo.

Para ver la definición de homología persistente con todo detalle se pueden consultar los libros [11] y [7].

Debido a la limitación de recursos computacionales, no era viable estudiar las características topológicas de todos los voxels etiquetados como lesión nueva en la máscara manual en el

dataset ImaginEM. Por ello, la estrategia consistió en escoger de forma aleatoria 20 casos que el modelo de 100 épocas etiquetara como lesiones nuevas que, comparándolos con la segmentación manual fueran TP, y otros 20 que, comparándolos con el ground truth fueran FP. Con el fin de calcular la homología persistente del bounding box de estos voxels para encontrar alguna relación que permitiera distinguir FP de TP. Para hallar la PH de los voxels, se recuperaron las imágenes FLAIR followup asociadas a estos FP y TP. Se estudió la PH sobre las imágenes FLAIR followup porque éstas tienen toda la información necesaria para hallar las diferencias topológicas que se encuentran entre una componente TP y FP. Cabe dejar constancia de que debido a las limitaciones computacionales, en vez de coger todo el volumen de la imagen FLAIR followup, se tomó el bounding box de cada TP y FP en coordenadas para extraer en la imagen FLAIR followup únicamente el volumen correspondiente a estas coordenadas y analizarlo mediante la PH en busca de diferencias. Se usó la imagen FLAIR followup, pero si se hubiera empleado por ejemplo la diferencia entre FLAIR followup y FLAIR baseline, se habría obtenido más información sobre el cambio producido en cada lesión, ayudando así a detectar avances en las lesiones. El coste computacional de esto es mayor, y las limitaciones computacionales no lo permitieron.

Escogidos los 20 FP y los 20 TP, el siguiente paso fue hallar el identificador de éstos para poder recuperar sus imágenes FLAIR followup (que van asociadas al ID). Se obtuvo que, los 20 FP correspondían a 15 IDs distintos, y los 20 TP correspondían a 13 IDs distintos. Es decir, las 20 componentes TP se repartían en 13 IDs distintos, y las 20 componentes FP en 15 IDs distintos, consiguiendo una distribución equilibrada. Una vez que se obtuvo para cada ID su imagen FLAIR followup, se calculó para cada subvolumen dos estadísticos, el número de “barras”, $n_{bars} h0$, que tenía cada componente y, el “tiempo de vida”, $sum\ life\ h0$, que tenía cada componente. Estos datos se agruparon por FP y TP, y lo que se obtuvo se refleja en la siguiente tabla:

Tipo	count	mean	std	min	25 %	50 %	75 %	max
FP	20.0	4.60	5.07	0.0	1.75	3.0	6.25	22.0
TP	20.0	22.55	38.75	0.0	4.00	7.0	16.50	139.0

Cuadro 4.17: Estadísticos descriptivos de $n_{bars_h0_fl}$ en FLAIR followup para falsos positivos (FP) y verdaderos positivos (TP).

Tipo	count	mean	std	min	25 %	50 %	75 %	max
FP	20.0	63.15	59.86	0.0	29.10	53.37	85.81	238.64
TP	20.0	231.92	368.99	0.0	52.39	81.03	178.91	1272.05

Cuadro 4.18: Estadísticos descriptivos de *sum_life_h0_fl* en FLAIR followup para falsos positivos (FP) y verdaderos positivos (TP).

En las tablas se observa una diferencia notable entre los valores de los TP y los FP. En cuanto al número de barras que se encuentra en cada componente, la clase de verdaderos positivos tiene una media casi cinco veces mayor a la de los falsos positivos. En cuanto a la “vida” de cada componente, la clase TP tiene una vida media casi cuatro veces mayor que los FP.

Representando gráficamente la distribución de estos estadísticos mediante gráficos de violín se obtiene:

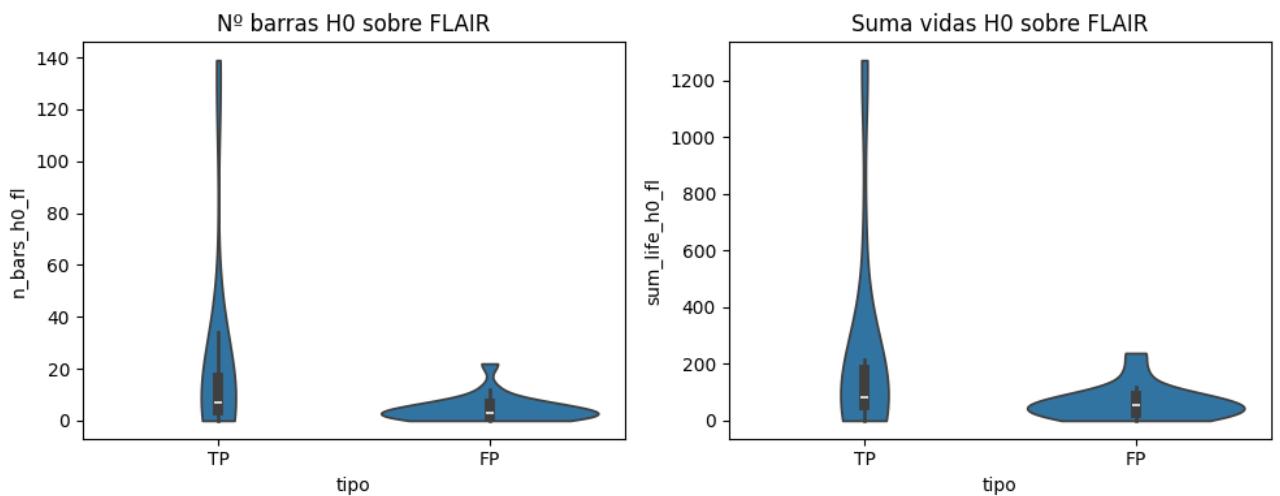


Figura 4.1: Gráficos violín estadísticos PH.

Respecto a los gráficos de violín: se confirma que efectivamente, los FP tienen muy pocas barras en comparación con los TP (los violines en los FP son estrechos mientras que el de TP es alargado, indicando que algunos TP tienen gran cantidad de barras/componentes en H_0). En cuanto a la suma de vidas ocurre algo análogo, en FP el violín se encuentra en un rango compacto, en comparación con el violín de TP, que se expande hasta 1200, indicando que la masa de topología H_0 (suma de vidas) es considerablemente mayor en TP.

Por tanto, se concluye que en el dataset ImaginEM sí se distingue topológicamente un FP y un TP. Para descartar en el postprocesado estos FP y mejorar las métricas, se abren dos vías, por una parte el número de barras, (introduciendo un filtro que pida que las lesiones etiquetadas como nuevas (etiqueta 2), deben tener $\geq k$ barras). Por otro lado, filtrando por la suma de vidas, (se toma un τ tal que la vida de cada voxel etiquetado con 2 debería sumar $\geq \tau$).

Para decidir qué estrategia tomar, si filtrar por número de barras (k) o por suma de vidas (τ), se cogieron varios k ($k=3, k=5, k=7, k=10$), y varios τ ($\tau=50, \tau=80, \tau=100, \tau=150$), y se analizó para qué valores se reducían más FP y se conservaban más TP. Obteniendo los siguientes resultados,

Parámetro	TP_keep	FP_keep
$k = 3$	18/20	12/20
$k = 5$	13/20	8/20
$k = 7$	11/20	5/20
$k = 10$	9/20	2/20
$\tau = 50$	15/20	11/20
$\tau = 80$	11/20	6/20
$\tau = 100$	9/20	3/20
$\tau = 150$	8/20	2/20

Cuadro 4.19: Fracción de verdaderos positivos (TP_keep) y falsos positivos (FP_keep) retenidos para distintos valores de k y del umbral τ .

Los mejores resultados se obtienen usando la estrategia de filtrar por el número de barras en H_0 , el que mejores cifras da es $k=3$, ya que casi mantiene la totalidad de los TP, y casi reduce a la mitad los FP. Debido a ello, en la estrategia de postprocesamiento por PH se usa de filtro el número de barras en H_0 , pidiendo que una lesión identificada por el modelo de 100 épocas como lesión nueva (etiqueta=2), debe tener ≥ 3 barras en H_0 , de no ser así, a los que fueron etiquetados por el modelo como lesión nueva, se les pasa a etiqueta 0.

Las métricas obtenidas al realizar este postprocesamiento basado en la homología persistente pidiendo que $k \geq 3$ en **test-split de ImaginEM** son:

3.1.2.1. Voxel-wise ImaginEM

Modelo	TP	FP	FN	TN	Dice new	Prec. new	Rec. new	F1 new
100_PH_K3	53878	7882	9237	518167	0.8629	0.8724	0.5528	0.6768

Cuadro 4.20: Resultados voxel-wise para la lesión nueva en ImaginEM (modelo 100_PH_K3).

Modelo	Prec. stable	Rec. stable	F1 stable	F1 macro	F1 weighted
100_PH_K3	0.9825	0.8223	0.8953	0.7860	0.8660

Cuadro 4.21: Resultados voxel-wise para la lesión estable y promedios en ImaginEM (modelo 100_PH_K3).

3.1.2.2. Lesion-wise ImaginEM

Modelo	GT lesiones	GT detectadas	FP lesiones	Sensibilidad	Precisión
100_PH_K3	389	230	51	0.5913	0.8185

Cuadro 4.22: Resultados lesion-wise en ImaginEM para el modelo 100_PH_K3.

3.1.2.3. ID-wise ImaginEM

Modelo	Casos	Casos con lesión nueva	Casos detectados	Sensibilidad
100_PH_K3	53	53	52	0.9811

Cuadro 4.23: Resultados ID-wise en ImaginEM para el modelo 100_PH_K3.

Las métricas obtenidas al realizar este postprocesamiento basado en la homología persistente pidiendo que $k \geq 3$ en **MSSEG2** son:

3.1.2.4. Voxel-wise MSSEG2

Modelo	TP	FP	FN	TN	Dice new	Prec. new	Rec. new	F1 new
100_PH_K3	34568	85838	25349	1322305445	0.3834	0.2871	0.5769	0.3834

Cuadro 4.24: Resultados voxel-wise para la lesión nueva en MSSEG2 (modelo 100_PH_K3).

Modelo	Prec. no_new	Rec. no_new	F1 no_new	F1 macro	F1 weighted
100_PH_K3	0.99998	0.99994	0.99996	0.6917	0.99993

Cuadro 4.25: Resultados voxel-wise para voxels sin lesión nueva y promedios en MSSEG2 (modelo 100_PH_K3).

3.1.2.5. Lesion-wise MSSEG2

Modelo	GT lesiones	GT detectadas	FP lesiones	Sensibilidad	Precisión
100_PH_K3	154	67	83	0.4351	0.4467

Cuadro 4.26: Resultados lesion-wise en MSSEG2 para el modelo 100_PH_K3.

3.1.2.6. ID-wise MSSEG2

Modelo	Casos	Casos con lesión nueva	Casos detectados	Sensibilidad
100_PH_K3	40	29	26	0.8966

Cuadro 4.27: Resultados ID-wise en MSSEG2 para el modelo 100_PH_K3.

3.1.3. Postprocesado usando PH+volumen (PH+P3)

En este postprocesado se aplicó el filtro de PH descrito en el punto anterior, y después se aplicó el filtro del percentil 3 del volumen.

3.1.4. Postprocesado usando volumen+PH (P3+PH)

En este postprocesado se aplicó el filtro del percentil 3 del volumen descrito en el primer punto, y después se aplicó el filtro de PH con $k \geq 3$.

Las métricas obtenidas tanto en PH+P3 como en P3+PH son iguales, esto es debido a que el filtro que se aplica usando la PH es menos agresivo que el percentil 3.

A continuación se muestran las métricas obtenidas en **test-split de ImaginEM**:

3.1.4.1. Voxel-wise ImaginEM

Modelo	TP	FP	FN	TN	Dice new	Prec. new	Rec. new	F1 new
100_K3_P3	51671	6844	9237	518167	0.8653	0.8830	0.5302	0.6626
100_P3_K3	51671	6844	9237	518167	0.8653	0.8830	0.5302	0.6626

Cuadro 4.28: Resultados voxel-wise para la lesión nueva en ImaginEM (modelos 100_K3_P3 y 100_P3_K3).

Modelo	Prec. stable	Rec. stable	F1 stable	F1 macro	F1 weighted
100_K3_P3	0.9825	0.8223	0.8953	0.7789	0.8641
100_P3_K3	0.9825	0.8223	0.8953	0.7789	0.8641

Cuadro 4.29: Resultados voxel-wise para la lesión estable y promedios en ImaginEM (modelos 100_K3_P3 y 100_P3_K3).

3.1.4.2. Lesion-wise ImaginEM

Modelo	GT lesiones	GT detectadas	FP lesiones	Sensibilidad	Precisión
100_K3_P3	389	169	16	0.4344	0.9135
100_P3_K3	389	169	16	0.4344	0.9135

Cuadro 4.30: Resultados lesion-wise en ImaginEM para los modelos 100_K3_P3 y 100_P3_K3.

3.1.4.3. ID-wise ImaginEM

Modelo	Casos	Casos con lesión nueva	Casos detectados	Sensibilidad
100_K3_P3	53	53	52	0.9811
100_P3_K3	53	53	52	0.9811

Cuadro 4.31: Resultados ID-wise en ImaginEM para los modelos 100_K3_P3 y 100_P3_K3.

Las métricas obtenidas en **MSSEG2** son:

3.1.4.4. Voxel-wise MSSEG2

Modelo	TP	FP	FN	TN	Dice new	Prec. new	Rec. new	F1 new
100_K3_P3	33305	81744	26612	1322309539	0.3807	0.2895	0.5559	0.3807
100_P3_K3	33305	81744	26612	1322309539	0.3807	0.2895	0.5559	0.3807

Cuadro 4.32: Resultados voxel-wise para la lesión nueva en MSSEG2 (modelos 100_K3_P3 y 100_P3_K3).

Modelo	Prec. no_new	Rec. no_new	F1 no_new	F1 macro	F1 weighted
100_K3_P3	0.99998	0.99994	0.99996	0.6903	0.99993
100_P3_K3	0.99998	0.99994	0.99996	0.6903	0.99993

Cuadro 4.33: Resultados voxel-wise para voxels sin lesión nueva y promedios en MSSEG2 (modelos 100_K3_P3 y 100_P3_K3).

3.1.4.5. Lesion-wise MSSEG2

Modelo	GT lesiones	GT detectadas	FP lesiones	Sensibilidad	Precisión
100_K3_P3	154	49	32	0.3182	0.6049
100_P3_K3	154	49	32	0.3182	0.6049

Cuadro 4.34: Resultados lesion-wise en MSSEG2 para los modelos 100_K3_P3 y 100_P3_K3.

3.1.4.6. ID-wise MSSEG2

Modelo	Casos	Casos con lesión nueva	Casos detectados	Sensibilidad
100_K3_P3	40	29	24	0.8276
100_P3_K3	40	29	24	0.8276

Cuadro 4.35: Resultados ID-wise en MSSEG2 para los modelos 100_K3_P3 y 100_P3_K3.

4. Conclusiones análisis cuantitativo de los distintos postprocesamientos

Evaluando las distintas modalidades de postprocesamiento, en el test-split de ImaginEM (lesión estable vs lesión nueva) se observa que todas las modalidades mantienen buenas métricas. El postprocesado basado en PH, con $k=3$, es el que mejor equilibrio presenta entre reducción de falsos positivos y mantenimiento de la sensibilidad y la precisión. Por otra parte, si se aplica únicamente postprocesado filtrando por volumen de las lesiones, se observa que percentiles agresivos como 5 o 10 mantienen prácticamente el mismo número de FP que un percentil 3, pero poseen una sensibilidad mucho menor, confirmando que estos filtros sacrifican casos clínicamente relevantes. El percentil 1 es muy tenue y casi no se diferencia de las métricas del modelo de 100 épocas, y es superado por el postprocesado del percentil 3 en volumen. El percentil 3 en volumen tiene buen equilibrio, pero no consigue superar la especificidad que aporta la homología (muy útil en la eliminación de falsos positivos, sin perder sensibilidad).

Analizando las métricas obtenidas en el dataset externo MSSEG2 (no lesión vs lesión nueva), los resultados obtenidos son mucho peores que los del conjunto anterior. Esto fue comentado anteriormente, se daba un desbalanceamiento importante en ImaginEM debido a que toda la población del dataset presentaba lesiones nuevas, lo que podía suponer que el modelo sufriera sobreajuste. Además, en la muestra de MSSEG2 formada por 40IDs, no todos ellos desarrollan lesiones nuevas. Esto se refleja en que en MSSEG2 la precisión para distinguir background es de casi el 100 % para todos, lo que indica que los errores se concentran en la detección de nuevas lesiones. Se puede observar que estos postprocesamientos apenas consiguen reducir los falsos positivos, esto puede ser debido a que en MSSEG2 sólo existen dos clases (0 background y 1 lesión nueva), mientras que en el conjunto ImaginEM se encuentran tres clases (0 background, 1 lesión estable, 2 lesión nueva). Además, como se comentó anteriormente, los voxels de las imágenes en MSSEG2 presentan hasta 33 combinaciones posibles en sus dimensiones, y algunas de ellas muy dispares entre sí, debido a que las imágenes por RM han sido tomadas en multitud de centros, con escáneres distintos... Los filtros de PH se calcularon a partir de las imágenes de ImaginEM, que tienen voxels con dimensiones muy similares, lo que hace que los filtros de PH hallados en ImaginEM no sean tan eficaces aplicados al dataset MSSEG2, donde la variabilidad es mucho mayor. Aún así, en MSSEG2 también se observa que el postprocesamiento que presenta mejor equilibrio entre la reducción de FP sin comprometer la sensibilidad ni la precisión, y mantener el mayor dice y F1, es la homología persistente, con $k=3$. Esta técnica ofrece un término medio, mejora la precisión respecto al modelo de 100 épocas sin degradar la sensibilidad como hacen por ejemplo los filtros percentil 5 y percentil 10 de volumen. En cuanto a la sensibilidad por caso, la homología presenta la tercera mayor (por detrás de 100-raw y P1),

por lo que se puede afirmar que la técnica topológica de filtración descarta FP sin perder tantos casos clínicos.

Por otra parte, se obtienen las mismas métricas realizando P3 (Percentil 3), P3+PH y PH+P3. Esto se debe a que la técnica de PH actúa sobre volúmenes inferiores a los que trata P3, por lo que independientemente del orden de aplicación, se obtienen los mismos resultados.

En conclusión, el uso de técnicas de análisis topológico, como en este caso la homología persistente, ayudan a eliminar FP, sin comprometer la sensibilidad, lo que es de gran importancia, pues de esta forma no se pierden IDs que tengan nuevas lesiones, de suma importancia a la hora de dar un diagnóstico.

5. Análisis cualitativo de los postprocesados en ImaginEM

El objetivo de esta sección es mostrar visualmente qué tan bien está detectando lesiones nuevas el modelo de 100 épocas con sus respectivos postprocesados en el dataset ImaginEM. Como se ha visto en la sección anterior, los modelos que han dado mejores métricas son 100-raw, PH K3 (el filtro de homología persistente con $k=3$) y P3 (únicamente filtro por volumen, Percentil 3). Por ello, se mostrarán las comparativas visuales únicamente para estos tres modelos. Para elegir sobre qué imágenes realizar esta comparativa, se escogió una imagen por cada uno de los cuatro criterios siguientes:

- Imagen en la que se obtuvieran las mejores métricas de predicción vs GT en lesiones nuevas
- Imagen en la que se obtuvieran las peores métricas de predicción vs GT en lesiones nuevas
- Imagen en la que se obtuvieran las mejores métricas de predicción vs GT en lesiones estables
- Imagen en la que se obtuvieran las peores métricas de predicción vs GT en lesiones estables

Para el primer criterio, se obtiene la imagen siguiente,

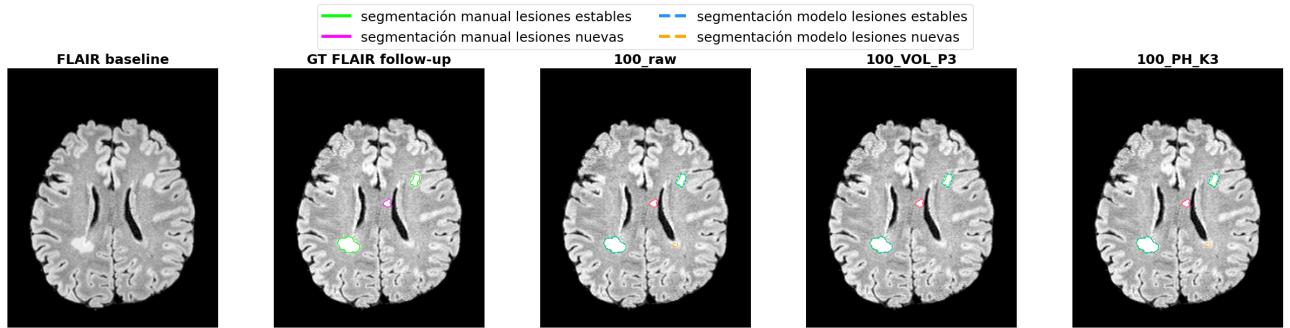


Figura 4.2: Segmentación manual vs segmentación modelo. En este caso las predicciones de lesiones nuevas son buenas.

Haciendo zoom en esta imagen se puede observar que en los modelos 100 raw y 100 PH K3 hay una zona segmentada como lesión nueva. Esto se corresponde con un falso positivo, ya que se comprueba que en la máscara manual no hay ninguna lesión detectada. En el modelo 100 VOL P3 esta zona no aparece segmentada como lesión nueva, puesto que su volumen no supera el filtro de P3. Esto se puede visualizar en la imagen siguiente, (se han cambiado los colores para que sea más fácil de distinguir cada tipo de segmentación).

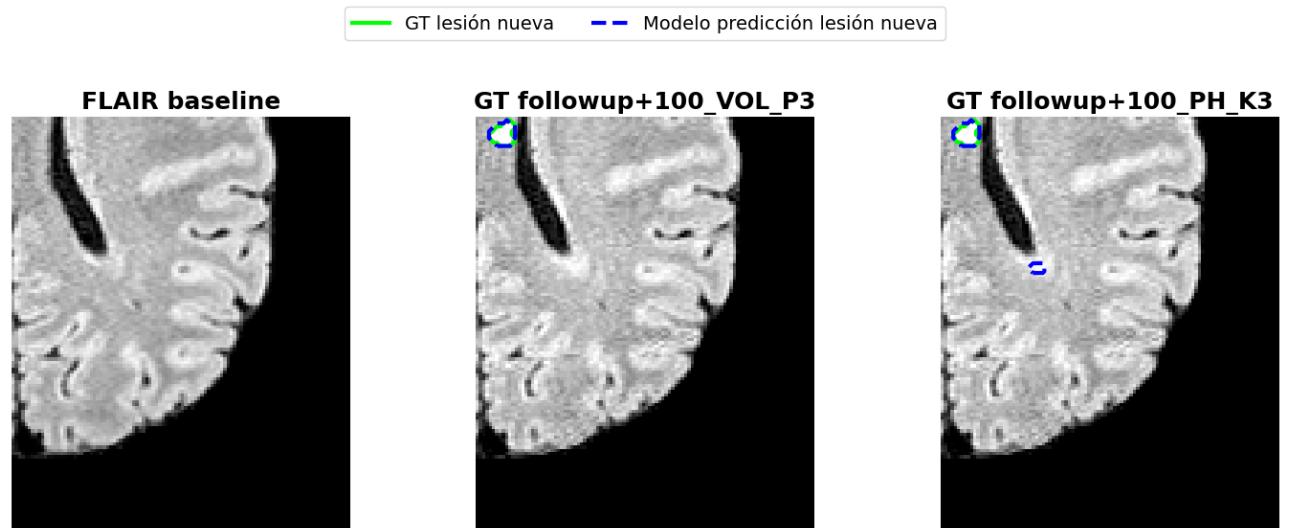


Figura 4.3: Segmentación 100 VOL P3 vs segmentación 100 PH K3.

Las métricas obtenidas en el modelo 100 PH K3 fueron,

Model	Prec. stable	Rec. stable	F1 stable	Prec. new	Rec. new	F1 new
100_PH_K3	1.0000	0.9142	0.9552	0.9036	0.9740	0.9375

Cuadro 4.36: Métricas voxel-wise por clase para el caso ImaginEM_MSVIS_095_03 con el modelo 100_PH_K3.

En la siguiente imagen se muestra un ejemplo de comparación de segmentaciones en las que los modelos tienen malas métricas en la predicción de lesiones nuevas,

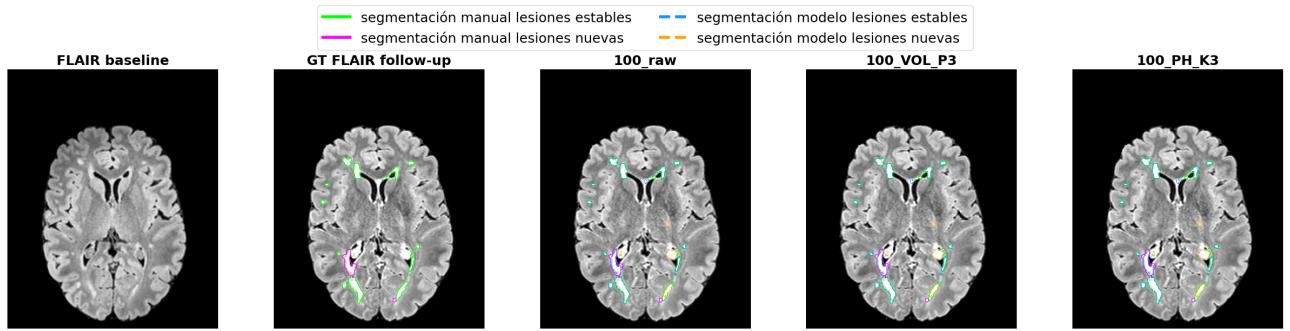


Figura 4.4: Segmentación manual vs segmentación modelo. En este caso las predicciones de lesiones nuevas son malas.

Si se hace zoom en la imagen, se comprueba que ninguna de las predicciones de lesiones nuevas producidas por los modelos coinciden con el GT. De hecho, en este caso, la mayor lesión nueva de la máscara manual, es etiquetada por los modelos como lesión estable. Esto mismo se ve reflejado en las métricas que se obtuvieron en el modelo 100 PH K3 para esta imagen,

Model	Prec. stable	Rec. stable	F1 stable	Prec. new	Rec. new	F1 new
100_PH_K3	0.9856	0.8803	0.9300	0.0	0.0	0.0

Cuadro 4.37: Métricas voxel-wise por clase para el caso ImaginEM_MSVIS_072_05 con el modelo 100_PH_K3.

En la siguiente imagen se muestra una comparación de segmentaciones en las que los modelos tienen buenas métricas en la predicción de lesiones estables,

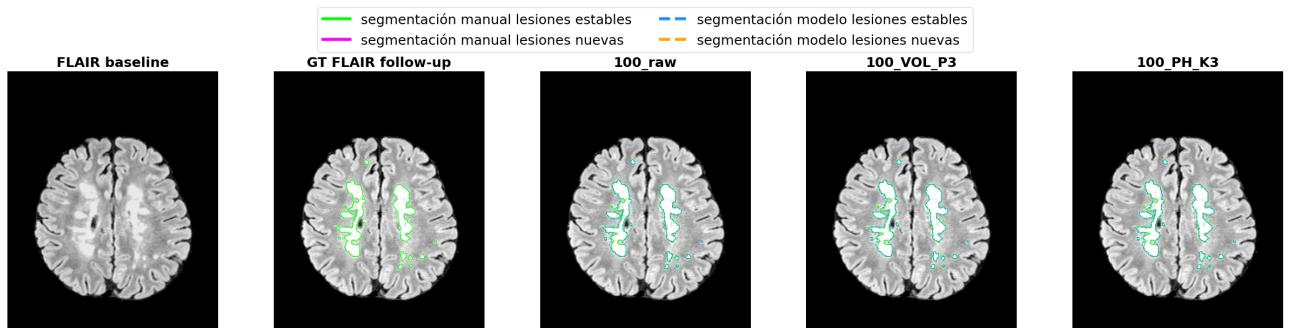


Figura 4.5: Segmentación manual vs segmentación modelo. En este caso las predicciones de lesiones estables son buenas.

Este caso en particular no desarrolla ninguna lesión nueva. Se observa que la delimitación de

las segmentaciones producidas por los modelos prácticamente coinciden con las de la máscara manual.

Las métricas obtenidas son las siguientes,

Model	Prec. stable	Rec. stable	F1 stable	Prec. new	Rec. new	F1 new
100_PH_K3	0.9999	0.9007	0.9477	0.4493	0.7884	0.5724

Cuadro 4.38: Métricas voxel-wise por clase para el caso ImaginEM_MSVIS_089_02 con el modelo 100_PH_K3.

En la siguiente imagen se muestra una comparación de segmentaciones en las que los modelos tienen malas métricas en la predicción de lesiones estables,

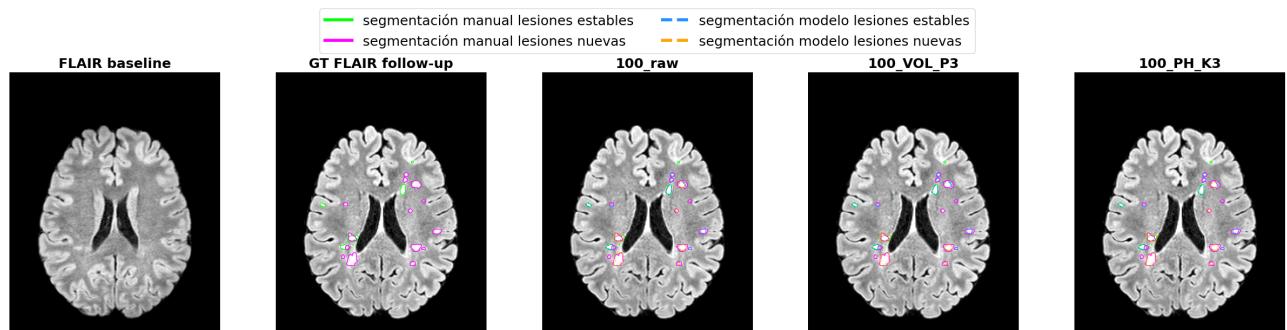


Figura 4.6: Segmentación manual vs segmentación modelo. En este caso las predicciones de lesiones estables son malas.

Mirando con detenimiento cada segmentación, se observa que hay varias zonas etiquetadas en la máscara manual como lesiones nuevas que los modelos han clasificado como lesiones estables, por ello se dan estas malas métricas, como se puede observar en la tabla siguiente,

Model	Prec. stable	Rec. stable	F1 stable	Prec. new	Rec. new	F1 new
100_PH_K3	0.3784	0.6683	0.4832	0.9901	0.3803	0.5495

Cuadro 4.39: Métricas voxel-wise por clase para el caso ImaginEM_MSVIS_088_05 con el modelo 100_PH_K3.

Con el fin de enseñar de un modo alternativo la forma en que realiza la segmentación cada modelo, en las imágenes siguientes se representa todo el área segmentada de cada predicción mostrando de distinto color los TP, FP y FN para los casos anteriormente seleccionados.

En esta imagen se observa cómo se aproxima la predicción de la lesión estable a la realidad. Los puntos susceptibles de tener FP o FN son aquellos que se encuentran en la frontera de la lesión.

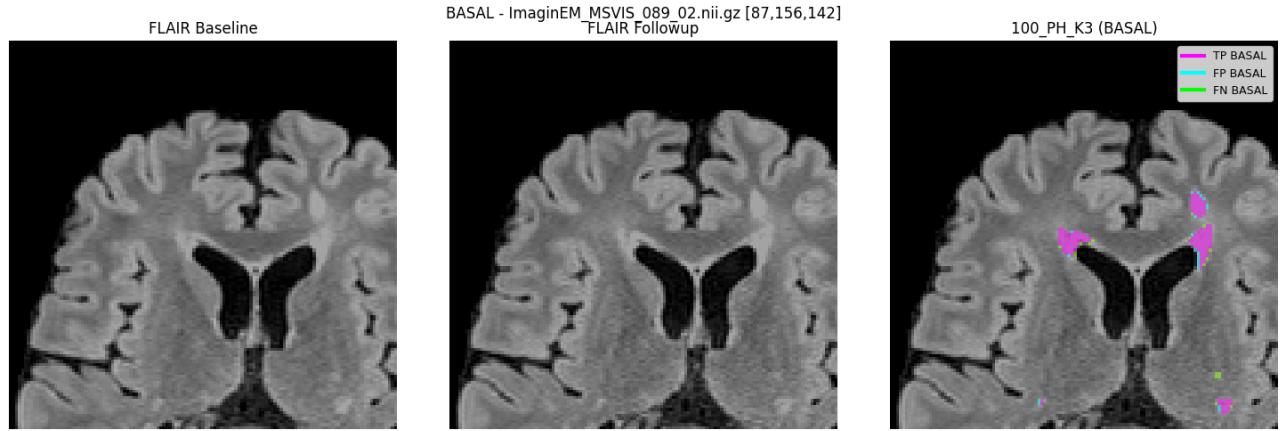


Figura 4.7: Comparativa TP, FP, FN en una predicción estable en ImaginEM.

En la siguiente imagen también se obtiene que los FP y los FN se dan en los puntos frontera de la lesión nueva.

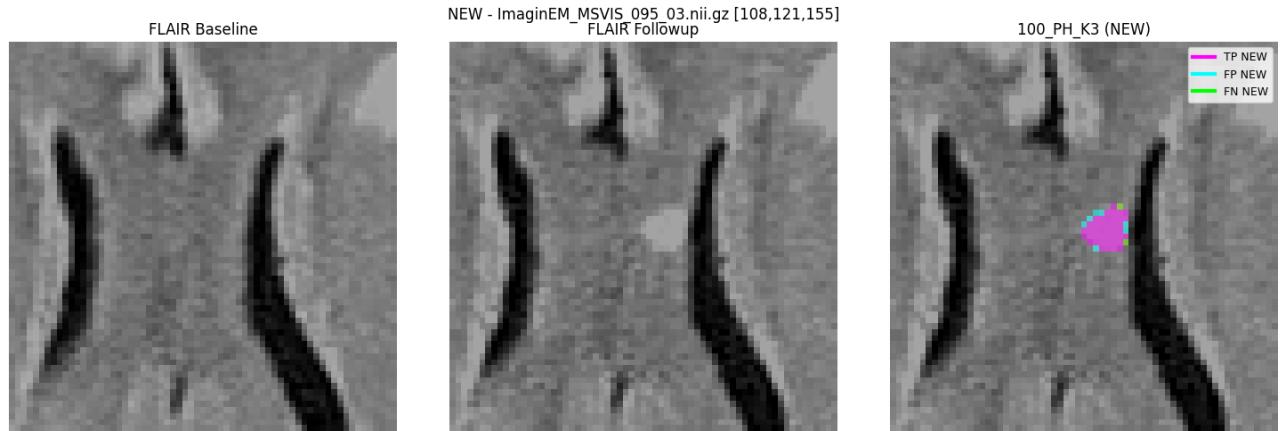


Figura 4.8: Comparativa TP, FP, FN en una predicción de lesión nueva en ImaginEM.

Lo que se acaba de ver es que los puntos susceptibles de sufrir FP o FN en las predicciones son aquellos que se encuentran en la frontera que delimita la zona sin lesión y la zona con lesión. Esto es debido a que las zonas limítrofes a la lesión a veces son similares a ésta y a la zona sin lesión.

6. Análisis cualitativo del postprocesado en MSSEG2

A continuación se muestra para una selección de cuatro imágenes de MSSEG2 la evolución de la lesión (FLAIR baseline vs FLAIR followup), junto con su segmentación manual y las predicciones que ha hecho el modelo 100 PH K3. En la imagen con el título TP/FP/FN NEW se compara la segmentación que hizo el modelo 100 PH K3 (etiquetas 0, 1, 2) con el GT de

la segmentación manual de esa imagen en MSSEG2 (etiqueta 1), pero de las predicciones de 100 PH K3 únicamente se toman aquellas que se clasificaron con etiqueta 2 (lesiones nuevas). En la última imagen de la serie se compara el GT de la segmentación manual de MSSEG2 con las predicciones que hizo el modelo 100 PH K3, tanto de lesiones estables (etiqueta 1), como de lesiones nuevas (etiqueta 2). Cabe destacar que debido a que MSSEG2 no posee la clase de lesión estable, no hay modo de comprobar si las lesiones estables que predijo el modelo 100 PH K3 son correctas o no.

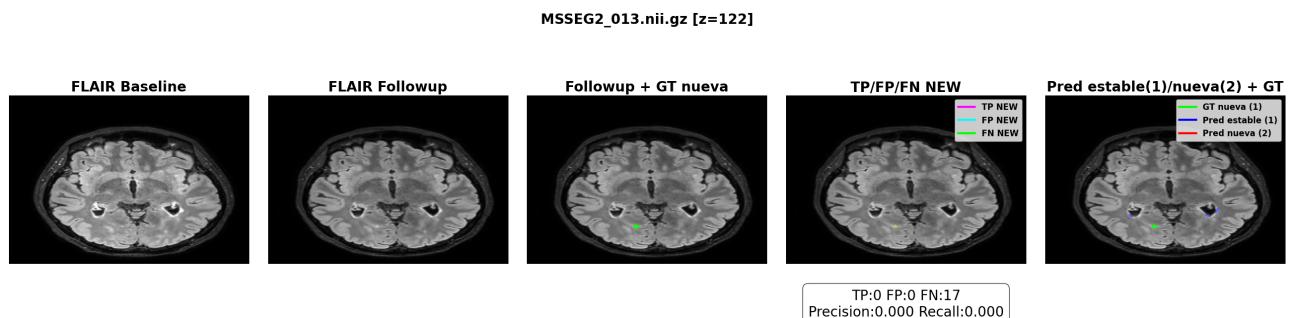


Figura 4.9: Evolución de la lesión en MSSEG2 y segmentación 100 PH K3.

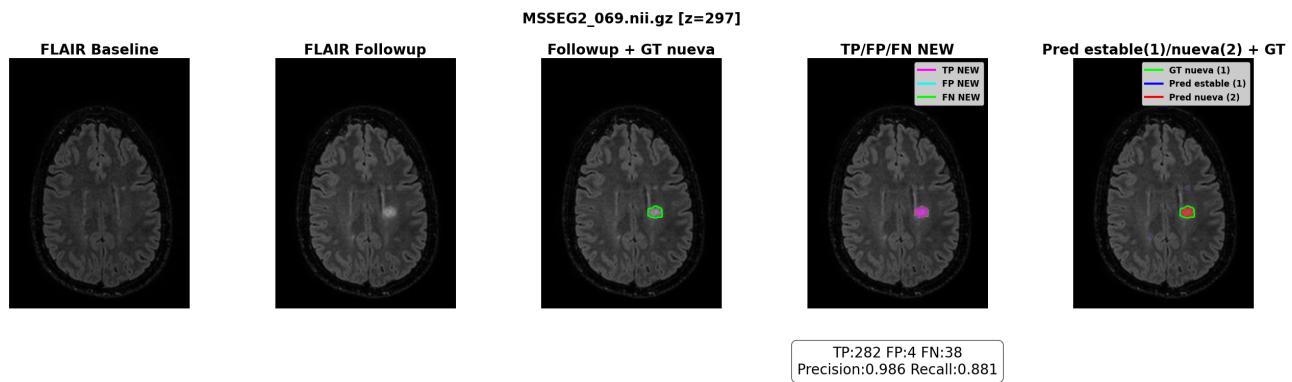


Figura 4.10: Evolución de la lesión en MSSEG2 y segmentación 100 PH K3.

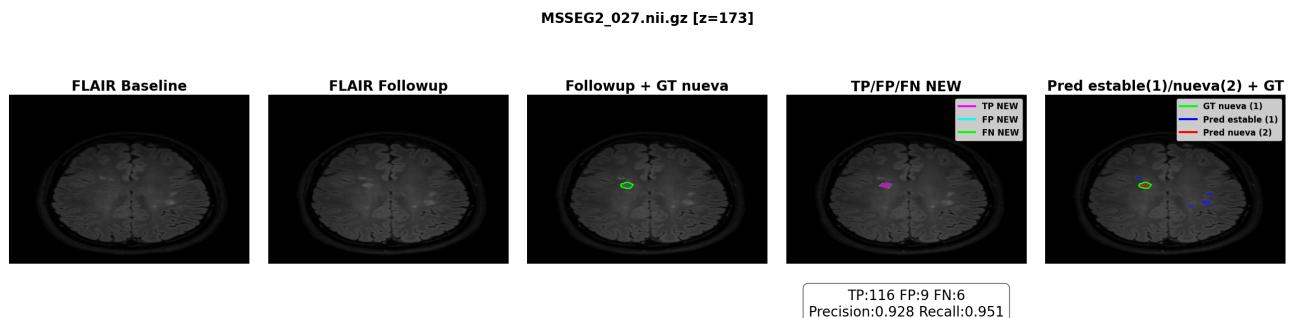


Figura 4.11: Evolución de la lesión en MSSEG2 y segmentación 100 PH K3.

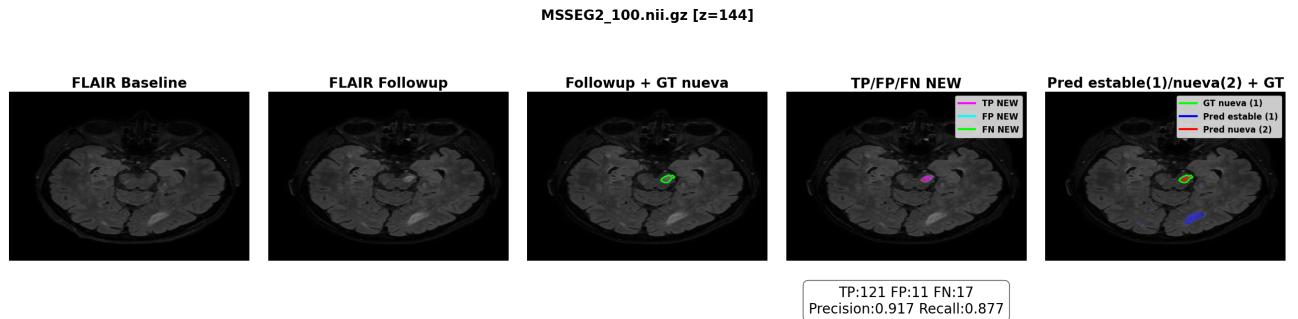


Figura 4.12: Evolución de la lesión en MSSEG2 y segmentación 100 PH K3.

En esta serie de imágenes se puede observar que, salvo el primer caso, el modelo 100 PH K3 es capaz de reconocer las lesiones nuevas, aunque, al igual que en el dataset ImaginEM, en los puntos frontera tiene dificultades y es ahí donde se encuentran FP y FN. En cuanto a las segmentaciones que ha hecho de lesiones estables, no es posible hallar sus métricas para valorarlas porque en la máscara de segmentación manual de MSSEG2 no existe esta clase. A pesar de ello, visualmente parece que el modelo es capaz de detectar las lesiones estables. Aún así, sería necesario que lo estudiara un radiólogo experto para extraer las conclusiones finales.

Capítulo 5

Conclusiones y trabajo futuro

En esta sección se va a hacer un repaso sobre los resultados obtenidos en este proyecto, las limitaciones con las que se ha tenido que lidiar, las áreas de mejora y los próximos pasos.

1. Conclusiones finales

Este proyecto ha consistido en desarrollar un modelo basado en la CNN nnU-Net v2. Los entrenamientos se han realizado a partir de un dataset (ImaginEM) formado por pares de imágenes RM longitudinales para cada uno de los IDs, con un total de 349 IDs con EM. Las imágenes se encuentran en secuencia FLAIR y fueron tomadas en dos tiempos distintos (FLAIR baseline y FLAIR followup, con una diferencia entre una y otra de 1 a 3 años). Todas ellas se tomaron en el Hospital Clínic de Barcelona y la máscara de segmentación manual contiene las etiquetas: 0 para zonas sin lesiones, 1 para zonas con lesiones estables y 2 para zonas con lesiones nuevas. Debido a las limitaciones computacionales ya comentadas anteriormente, se hicieron cuatro entrenamientos distintos de este modelo, tomando 20 épocas, 50 épocas, 100 épocas y 250 épocas, viendo que el que aportaba mejores métricas tanto en el test-split de ImaginEM, como en el dataset externo MSSEG2 era el de 100 épocas.

El elemento innovador introducido en este proyecto ha sido la aplicación de la homología persistente en la etapa de postprocesamiento, con el fin de reducir falsos positivos, pero sin comprometer la sensibilidad general del modelo. Con este enfoque topológico se consigue capturar invariantes geométricos asociados a las zonas lesionadas, como son: el número de componentes conexas o su persistencia temporal. Con estos invariantes se puede diferenciar ruido de las lesiones reales, ya que el ruido tiene una persistencia temporal inferior a las lesiones reales. Esta técnica ha demostrado ser más efectiva que los filtros por volumen de lesión tradicionales. Tanto en ImaginEM, como en MSSEG2, la PH permite obtener un mejor equilibrio entre FP, DICE, F1 y sensibilidad. Cabe puntualizar que los mejores resultados se obtienen en el dataset

ImaginEM, mientras que en el dataset MSSEG2 no se alcanza el 39 % en DICE score. Esto es debido a que las imágenes de ImaginEM fueron tomadas en un único centro (Hospital Clínic de Barcelona), con escáneres distintos a los del dataset MSSEG2 y, probablemente protocolos distintos, como se observó en el estudio de las dimensiones de los voxels de las imágenes de cada dataset. ImaginEM únicamente tenía 3 combinaciones distintas de las dimensiones de los voxels, similares entre sí. Mientras que, en MSSEG2 se observaron 33 variaciones distintas de dimensiones, algunas muy dispares entre sí. Debido a esto, los filtros de PH que se aplicaron en ImaginEM funcionaron bien. En MSSEG2 no tuvieron un efecto tan notable porque la PH se halló sobre los voxels de ImaginEM, similares entre sí, y al aplicarlos sobre los voxels de MSSEG2, que presentan gran variabilidad, esto no fue extensible. Aún así, los resultados demuestran que el desarrollo de modelos basados en la nnU-Net v2, junto con el uso de técnicas de análisis de datos topológico, (como PH), es una herramienta muy poderosa para aliviar la carga diagnóstica de los radiólogos en entornos de segmentación médica no estandarizada.

2. Limitaciones

El proyecto se ha visto afectado por varias limitaciones computacionales como, por ejemplo, ausencia de servidor/máquina virtual/tarjeta gráfica dedicada con RAM superior a 12GB, lo que ha restringido el alcance y rendimiento potencial del proyecto. Además, por ausencia de financiación no se pudieron usar aceleradores en la nube como Azure o AWS. La única opción viable fue realizar una suscripción a Google Colab Pro+. Ésta permitía el uso de tarjetas gráficas A100 de uso compartido con aproximadamente 16/17.5GB de memoria, con la limitación añadida de que las sesiones de Google Colab Pro+ duran un máximo de 24 horas. El entrenamiento en un solo fold en la nnU-Net v2 con 1000 épocas requería 40 horas de ejecución, lo que imposibilitó que se realizara cross validation, y seguía sin ser viable realizar un entrenamiento con 1000 épocas en un solo fold. En consecuencia, se vio la obligación de restringir los entrenamientos a configuraciones de 20, 50, 100 y 250 épocas.

Esta limitación de computación también afectó al postprocesamiento basado en la homología persistente, ya que no se pudo aplicar a todos los voxels predichos como lesiones nuevas por el modelo en los 349 IDs. En su lugar, ha sido necesario optar por hacer el estudio de la PH en una muestra estratificada obtenida a partir de 20 FP aleatorios y 20 TP. Además, en lugar de hallar la PH en cada uno de estos voxels, se tuvo que hallar la PH dentro del bounding box de cada uno de estos en su imagen FLAIR followup. Habría sido más fructífero haber hallado las características topológicas de las lesiones nuevas y basales sobre la diferencia entre FLAIR followup y FLAIR baseline de los 349 IDs. Por otro lado, únicamente se pudo hallar la homología de dimensión 0 (H_0), pero habría sido igual de enriquecedor haber hallado también

H_1 y H_2 , especialmente teniendo en cuenta que, como se concluye en [16], las curvas de Betti de dimensión 1 y 2 son las que mejor capturan las diferencias que pueda haber entre estructuras.

3. Áreas de mejora

En primer lugar, sería fundamental superar la limitación computacional. Esta mejora permitiría ejecutar un entrenamiento con cross validation, iniciando con un learning rate cercano a 0.001, inferior al que viene preestablecido en el entrenamiento en la nnU-Net v2 (0.01), y añadiendo un early stopping. Esta técnica estabilizaría la convergencia y mitigaría el sobreajuste que por ejemplo se ha dado en el entrenamiento realizado con 250 épocas (donde el DICE es menor que en el de 100 épocas). Otra mejora fundamental puede ser la extracción, durante el entrenamiento, de las características topológicas de cada voxel, tales como número de barras en H_0 , suma de vidas... La obtención de estos invariantes se haría a partir de las imágenes FLAIR baseline y FLAIR followup, con el fin de poder recoger información sobre la evolución de cada lesión y, además, detectar nuevas lesiones de forma precisa. Esta técnica superaría al postprocesamiento en homología persistente que se ha realizado en este proyecto, siendo esperable que se reduzca el número de FP, ya que de partida habría muchas zonas correspondientes a ruido que no se considerarían como positivos (lesiones nuevas). Como se comentó en el apartado de limitaciones, sería deseable hallar, además de las características topológicas de H_0 (componentes conexas), las de dimensiones superiores H_1 y H_2 con el fin de recoger más información que pueda mostrar las diferencias entre una zona lesionada y una zona sin lesión. Adicionalmente, una vez superadas las limitaciones de computación, se podría mejorar el postprocesado con objeto de buscar patrones que pudieran indicar las diferencias topológicas entre una lesión real y un FP. Para ello, sería deseable hallar los invariantes topológicos de todas las lesiones predichas como nuevas por el modelo en la diferencia FLAIR followup-FLAIR baseline (en lugar de sólo en FLAIR followup), y compararlas con su GT. También, se ha observado que en ImaginEM los voxels tienen dimensiones muy similares, esto hace que el modelo entrenado pueda verse sesgado, ya que la muestra no es representativa al no tener variabilidad. Se ha podido asegurar que otros datasets, como por ejemplo MSSEG2, poseen gran variabilidad de tipos de voxels, (debido a que han tomado imágenes por RM con distintos escáneres). Para que el modelo entrenado tenga capacidad de generalización, sin verse influenciado por una muestra que no tiene variabilidad, habría que añadir imágenes que tuvieran distintos tipos de voxels. Con ello se podría obtener una muestra representativa y balanceada, con el fin de que el modelo pudiera recoger todas las características, incluyendo las topológicas, recogidas con técnicas de PH. Esto se traduciría en que la evaluación externa del modelo entrenado en datasets independientes, como por ejemplo MSSEG2, tendría mejores métricas. Por último, otro

procedimiento que podría aportar más información sería añadir al uso de la secuencia FLAIR la implementación de las secuencias T1 y T2, de forma que se obtuviera más información sobre la estructura de cada lesión.

4. Próximos pasos

Los resultados obtenidos indican que con el análisis topológico de datos (TDA), en concreto mediante homología persistente, es posible hallar diferencias claras entre los TP y los FP a la hora de segmentar imágenes para detectar lesiones. El número de FP disminuye considerablemente en ImaginEM, sin comprometer la sensibilidad. Sin embargo, es preceptivo destacar que estas características de homología persistente se ven sesgadas debido a que las dimensiones de los voxels en ImagenEM son muy similares, dando lugar a que en MSSEG2 la aplicación de los filtros seleccionados a partir de ImaginEM no den una reducción de FP tan notable. En los siguientes proyectos habría que centrar la investigación en la homología persistente, con la financiación debida para poder implementar sin restricciones. Con el fin de evitar sesgos causados por los escáneres con los que se ha tomado cada imagen, a este desarrollo de modelos basados en la nnU-Net v2 junto con el estudio de las estructuras topológicas, sería conveniente añadir el entrenamiento de un modelo en una muestra representativa de imágenes con distintas dimensiones de voxels (como por ejemplo el dataset MSSEG2). Otro método a considerar para evitar dichos sesgos consistiría en la incorporación de un federated learning entre centros hospitalarios de todo el mundo, para así entrenar los modelos con la mayor variabilidad posible y conseguir predicciones más fiables que sirvieran de apoyo a los radiólogos especialistas para el diagnóstico y seguimiento de la enfermedad.

Bibliografía

- [1] Nadezhda Alsaanova, Pavel Bartenev, Maksim Sharaev, Milos Ljubisavljevic, Taleb Al Mansoori, and Yauhen Statsenko. Integrating Radiomics with Deep Learning Enhances Multiple Sclerosis Lesion Delineation, June 2025. arXiv:2506.14524 [eess].
- [2] Lei Bai, Dongang Wang, Hengrui Wang, Michael Barnett, Mariano Cabezas, Weidong Cai, Fernando Calamante, Kain Kyle, Dongnan Liu, Linda Ly, Aria Nguyen, Chun-Chien Shieh, Ryan Sullivan, Geng Zhan, Wanli Ouyang, and Chenyu Wang. Improving multiple sclerosis lesion segmentation across clinical sites: A federated learning approach with noise-resilient training. *Artificial Intelligence in Medicine*, 152:102872, June 2024.
- [3] Berke Doga Basaran, Paul M. Matthews, and Wenjia Bai. New lesion segmentation for multiple sclerosis brain images with imaging and lesion-aware augmentation. *Front. Neurosci.*, 16:1007453, October 2022.
- [4] Priyanka Belwal and Surendra Singh. Deep Learning techniques to detect and analysis of multiple sclerosis through MRI: A systematic literature review. *Computers in Biology and Medicine*, 185:109530, February 2025.
- [5] Alessandro Pasquale De Rosa, Marco Benedetto, Stefano Tagliaferri, Francesco Bardozzo, Alessandro D'Ambrosio, Alvino Biscecco, Antonio Gallo, Mario Cirillo, Roberto Tagliaferri, and Fabrizio Esposito. Consensus of algorithms for lesion segmentation in brain MRI studies of multiple sclerosis. *Sci Rep*, 14(1):21348, September 2024.
- [6] Marcos Diaz-Hurtado, Eloy Martínez-Heras, Elisabeth Solana, Jordi Casas-Roma, Sara Llufrí, Baris Kanber, and Ferran Prados. Recent advances in the longitudinal segmentation of multiple sclerosis lesions on magnetic resonance imaging: a review. *Neuroradiology*, 64(11):2103–2117, November 2022.
- [7] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. 01 2010.
- [8] Ander Elkoroaristizabal, Francesc Vivó, Albert Calvi, Elisabeth Solana, Elisabet Lopez-Soley, Salut Alba-Arbalat, Marcos Diaz-Hurtado, Baris Kanber, Jordi Casas-Roma, Sara

- Llufriu, Ferran Prados, and Eloy Martínez-Heras. Multiclass Lesion Detection Using Longitudinal MRI in Multiple Sclerosis. In Teresa Alsinet, Xavier Vilasís, Daniel García, and Elena Álvarez, editors, *Frontiers in Artificial Intelligence and Applications*. IOS Press, September 2024.
- [9] Nils Gessert, Julia Krüger, Roland Opfer, Ann-Christin Ostwaldt, Praveena Manogaran, Hagen H. Kitzler, Sven Schippling, and Alexander Schlaefer. Multiple sclerosis lesion activity segmentation with attention-guided two-path CNNs. *Computerized Medical Imaging and Graphics*, 84:101772, September 2020.
- [10] Francesco Guarnera, Alessia Rondinella, Elena Crispino, Giulia Russo, Clara Di Lorenzo, Davide Maimone, Francesco Pappalardo, and Sebastiano Battiatto. MSLesSeg: baseline and benchmarking of a new Multiple Sclerosis Lesion Segmentation dataset. *Sci Data*, 12(1):920, May 2025.
- [11] Allen Hatcher. *Algebraic topology*. Cambridge university press, New York, 2001.
- [12] Sarah Hindawi, Bartłomiej Szubstarski, Eric Boernert, Björn Tackenberg, and Jens Wuerfel. Federated learning for lesion segmentation in multiple sclerosis: a real-world multi-center feasibility study. *Front. Neurol.*, 16:1620469, September 2025.
- [13] Xiaoling Hu, Annabel Sorby-Adams, Frederik Barkhof, W. Taylor Kimberly, Oula Puonti, and Juan Eugenio Iglesias. P-Count: Persistence-based Counting of White Matter Hypointensities in Brain MRI, March 2024. arXiv:2403.13996 [eess].
- [14] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*, 18(2):203–211, February 2021.
- [15] Reda Abdellah Kamraoui, Boris Mansencal, José V. Manjon, and Pierrick Coupé. Longitudinal detection of new MS lesions using deep learning. *Front. Neuroimaging*, 1:948235, August 2022.
- [16] Toni Lozano-Bagén, Eloy Martinez-Heras, Giuseppe Pontillo, Elisabeth Solana, Francesc Vivó, Maria Petracca, Alberto Calvi, Sandra Garrido-Romero, Albert Solé-Ribalta, Sara Llufriu, Ferran Prados, and Jordi Casas-Roma. Evaluating topological and graph-theoretical approaches to extract complex multimodal brain connectivity patterns in multiple sclerosis. *Health Inf Sci Syst*, 13(1):68, October 2025.

- [17] Filip Orzan, Ştefania D. Iancu, Laura Dioşan, and Zoltán Bálint. Textural analysis and artificial intelligence as decision support tools in the diagnosis of multiple sclerosis – a systematic review. *Front. Neurosci.*, 18:1457420, January 2025.
- [18] Andreas Pommert, Ulf Tiede, and Karl Heinz Höhne. Volume Visualization. In *Brain Mapping: The Methods*, pages 707–723. Elsevier, 2002.
- [19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, May 2015. arXiv:1505.04597 [cs].
- [20] Àlex Rovira, Cristina Auger, and Juli Alonso. Magnetic resonance monitoring of lesion evolution in multiple sclerosis. *Ther Adv Neurol Disord*, 6(5):298–310, September 2013.
- [21] Melike Sah and Cem Direkoglu. A survey of deep learning methods for multiple sclerosis identification using brain mri images. *Neural Computing and Applications*, 34:1–25, 05 2022.
- [22] Melike Sah and Cem Direkoglu. A survey of deep learning methods for multiple sclerosis identification using brain MRI images. *Neural Comput & Applic*, 34(10):7349–7373, May 2022.
- [23] Peyman Tahghighi, Yunyan Zhang, Roberto Souza, and Amin Komeili. Enhancing New Multiple Sclerosis Lesion Segmentation via Self-supervised Pre-training and Synthetic Lesion Integration. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15008, pages 263–272. Springer Nature Switzerland, Cham, 2024. Series Title: Lecture Notes in Computer Science.
- [24] Sabina Umirzakova, Muksimova Shakhnoza, Mardieva Sevara, and Taeg Keun Whangbo. Deep learning for multiple sclerosis lesion classification and stratification using MRI. *Computers in Biology and Medicine*, 192:110078, June 2025.
- [25] Sergi Valverde, Mariano Cabezas, Eloy Roura, Sandra González-Villà, Deborah Pareto, Joan C. Vilanova, Lluís Ramió-Torrentà, Àlex Rovira, Arnau Oliver, and Xavier Lladó. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *NeuroImage*, 155:159–168, July 2017.
- [26] Stephen G. Wahlig, Pierre Nedelec, David A. Weiss, Jeffrey D. Rudie, Leo P. Sugrue, and Andreas M. Rauschecker. 3D U-Net for automated detection of multiple sclerosis lesions: utility of transfer learning from other pathologies. *Front. Neurosci.*, 17:1188336, October 2023.

- [27] Tun Wiltgen, Julian McGinnis, Sarah Schlaeger, Florian Kofler, CuiCi Voon, Achim Berthele, Daria Bischl, Lioba Grundl, Nikolaus Will, Marie Metz, David Schinz, Dominik Sepp, Philipp Prucker, Benita Schmitz-Koep, Claus Zimmer, Bjoern Menze, Daniel Rueckert, Bernhard Hemmer, Jan Kirschke, Mark Mühlau, and Benedikt Wiestler. LST-AI: A deep learning ensemble for accurate MS lesion segmentation. *NeuroImage: Clinical*, 42:103611, 2024.
- [28] Jin Ye, Son Duy Dao, Yicheng Wu, Yasmeen George, Daniel F Schmidt, Hengcan Shi, Winston Chong, and Jianfei Cai. New Multiple Sclerosis Lesion Segmentation via Calibrated Inter-patch Blending.
- [29] Yi Zhu, Thomas Grenier, and Chantal Revol-Muller. Comparative analysis of three advanced deep learning algorithms for Multiple Sclerosis lesion segmentation in FLAIR MRI. In *2024 IEEE 17th International Conference on Signal Processing (ICSP)*, pages 697–702, Suzhou, China, October 2024. IEEE.