UNIVERSIDAD DE MURCIA

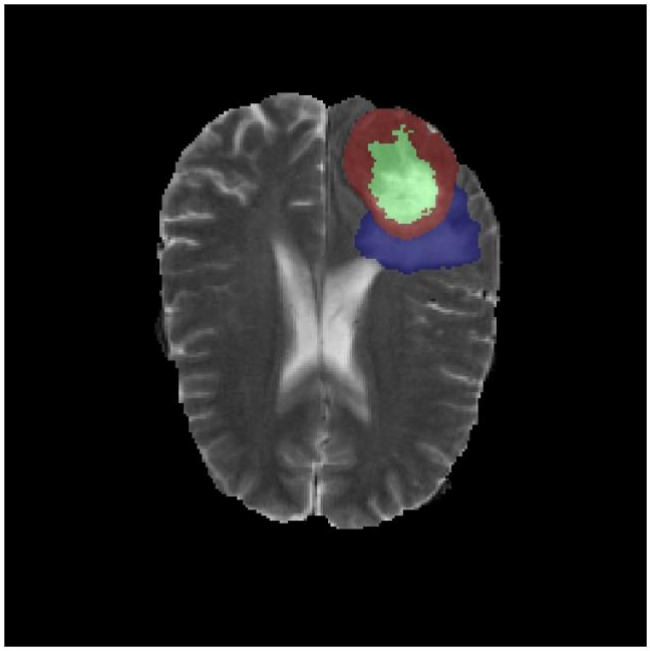
Facultad de Biología
Universidad de Murcia

# TRABAJO FINAL ANÁLISIS DE BIOIMÁGENES

## Máster en Bioinformática

## Facultad de Biología – Universidad de Murcia

# Segmentación automática de tumores cerebrales en resonancia magnética multimodal mediante nnU-Net



**Autores:** Antonio Gutiérrez Alcantud, Alejandro Antonio Rodríguez Garrido, Irene Xue-Feng Lope Mateo

**Curso 2025-2026**

# ÍNDICE

# 1. RESUMEN

La segmentación automática de tumores cerebrales en imágenes de resonancia magnética (MRI) constituye una tarea fundamental en el diagnóstico asistido por ordenador y en la planificación del tratamiento oncológico. En este trabajo se presenta el desarrollo y la evaluación de un modelo de segmentación basado en **nnU-Net**, entrenado sobre imágenes MRI multimodales procedentes del dataset público BraTS. El modelo se entrenó utilizando volúmenes tridimensionales multicanal y se evaluó sobre un conjunto de prueba independiente mediante métricas cuantitativas de solapamiento, estimación volumétrica y precisión en los bordes, así como mediante un análisis cualitativo de las segmentaciones generadas. Los resultados obtenidos muestran un alto grado de concordancia entre las segmentaciones automáticas y las anotaciones manuales de referencia, con un rendimiento consistente en la mayoría de los casos evaluados. Asimismo, el análisis visual confirma la coherencia anatómica de las segmentaciones y permite identificar las principales limitaciones del modelo en regiones tumorales complejas. En conjunto, este estudio pone de manifiesto el potencial de nnU-Net como herramienta robusta para la segmentación automática de tumores cerebrales en MRI multimodal, contribuyendo al avance de métodos reproducibles y eficientes en el ámbito del análisis de imagen médica.

# 2. INTRODUCCIÓN

## 2.1. Importancia de la segmentación de tumores cerebrales

La segmentación de tumores cerebrales en imágenes de resonancia magnética (MRI) constituye una etapa crítica dentro del proceso de diagnóstico asistido por computador, ya que permite localizar y delimitar de forma precisa las distintas regiones tumorales presentes en el cerebro. El diagnóstico computacional de tumores cerebrales se compone fundamentalmente de tres fases: detección, segmentación y clasificación, siendo la segmentación la encargada de aislar los tejidos tumorales de los tejidos sanos circundantes (Abd-Ellah *et al.* , 2019; Bauer *et al.* , 2013; Menze *et al.* , 2015).

La correcta segmentación del tumor proporciona información clínica esencial, como la localización, el tamaño, la forma y la extensión de las diferentes subregiones tumorales, incluyendo el núcleo tumoral, el edema peritumoral y las áreas necróticas. Esta información resulta clave para la planificación del tratamiento, la monitorización de la evolución de la enfermedad y la evaluación de la respuesta terapéutica, especialmente en tumores de tipo glioma, que presentan una elevada heterogeneidad estructural y biológica (Menze *et al.* , 2015; Havaei *et al.* , 2017; Abd-Ellah *et al.* , 2019).

Tradicionalmente, la segmentación de tumores cerebrales se ha realizado de forma manual por radiólogos expertos. Sin embargo, este procedimiento consume un tiempo considerable y presenta una elevada variabilidad dependiendo del observador, especialmente en tumores con bordes difusos o regiones con bajo contraste en las imágenes MRI. Estas limitaciones han motivado el desarrollo de métodos automáticos y

semiautomáticos que permitan mejorar la precisión, reproducibilidad y eficiencia del proceso de segmentación (Abd-Ellah *et al.* , 2019; Pereira *et al.* , 2016).

## 2.2. Resonancia magnética y Deep Learning en el análisis de tumores cerebrales

La resonancia magnética (MRI) es la modalidad de imagen más empleada en el diagnóstico de tumores cerebrales debido a su excelente contraste entre tejidos blandos y su capacidad para representar diferentes propiedades anatómicas y fisiológicas del cerebro, como la localización, morfología y extensión de las lesiones tumorales, mediante múltiples secuencias de adquisición (Bauer *et al.* , 2013; Menze *et al.* , 2015; Abd-Ellah *et al.* , 2019).

No obstante, la complejidad de las imágenes MRI, unida a la gran variabilidad en la apariencia de los tumores cerebrales, dificulta el análisis manual exhaustivo de los estudios radiológicos. En este contexto, el uso de técnicas de Deep Learning ha emergido como una solución eficaz para automatizar tareas de análisis de imagen médica, especialmente aquellas relacionadas con la segmentación y caracterización de tumores. Los métodos basados en redes neuronales profundas superan a los enfoques tradicionales al ser capaces de aprender representaciones jerárquicas directamente a partir de los datos, sin necesidad de definir manualmente características específicas (Havaei *et al.* , 2017; Ottom *et al.* , 2022).

En particular, las redes neuronales convolucionales (CNN) han demostrado una elevada capacidad para modelar patrones espaciales complejos presentes en imágenes MRI, lo que resulta especialmente relevante en tumores cerebrales con bordes irregulares y regiones heterogéneas. Los modelos de *Deep Learning* permiten integrar información procedente de distintas secuencias MRI, mejorando la robustez y precisión de los sistemas de análisis automático frente a variaciones en el contraste y el ruido de la imagen (Havaei *et al.* , 2017; Pereira *et al.* , 2016). Además, la aplicación de *Deep Learning* en imágenes MRI no solo contribuye a mejorar la precisión diagnóstica, sino que también reduce la carga de trabajo de los especialistas y favorece la reproducibilidad de los resultados (Ottom *et al.* , 2022; Isensee *et al.* , 2021).

## 2.3. Justificación del uso de U-Net y nnU-Net

Las arquitecturas basadas en U-Net se han consolidado como una de las soluciones más eficaces para tareas de segmentación en imágenes médicas, especialmente en el análisis de imágenes de resonancia magnética. U-Net fue diseñada específicamente para segmentación biomédica, combinando un esquema encoder–decoder con conexiones de salto (*skip connections*) que permiten preservar información espacial de alta resolución durante el proceso de reconstrucción de la imagen segmentada (Munir *et al.* , 2022).

Estas conexiones de salto constituyen una de las principales ventajas de U-Net frente a arquitecturas convolucionales convencionales, ya que facilitan la propagación del gradiente durante el entrenamiento y permiten integrar simultáneamente información

contextual global y detalles locales. Esto resulta especialmente relevante en la segmentación de tumores cerebrales, donde las regiones patológicas presentan formas irregulares, tamaños variables y bordes poco definidos. U-Net logra un equilibrio adecuado entre precisión y eficiencia computacional, incluso cuando se dispone de conjuntos de datos relativamente limitados (Munir *et al.* , 2022; Havaei *et al.* , 2017).

No obstante, el diseño manual de la arquitectura y la selección óptima de hiperparámetros continúan siendo aspectos críticos que pueden influir significativamente en el rendimiento del modelo. Por ello, nnU-Net surge como una extensión del enfoque U-Net tradicional, adaptando la arquitectura, los parámetros de entrenamiento y las estrategias de preprocesado en función de las características específicas del conjunto de datos, reduciendo la necesidad de intervención manual, y mejorando así la reproducibilidad de los resultados (Isensee *et al.* , 2021).

# 3. Materiales y métodos

## 3.1. Dataset

En este trabajo se empleó un conjunto de imágenes de resonancia magnética (MRI) de tumores cerebrales procedentes del dataset público **BraTS** (*Brain Tumor Segmentation Challenge*)**,** ampliamente utilizado como referencia en tareas de segmentación biomédica y evaluación de algoritmos de aprendizaje profundo.

Cada caso del dataset se almacena como un archivo NIfTI comprimido (.nii.gz), que representa el cerebro como un volumen tridimensional compuesto por una rejilla regular de elementos espaciales denominados vóxeles. Cada vóxel corresponde a una pequeña porción cúbica del tejido cerebral y constituye la unidad básica de análisis en los algoritmos de segmentación automática. En tareas de segmentación de tumores cerebrales, el objetivo consiste en asignar a cada vóxel una etiqueta que indica su pertenencia a una región anatómica o patológica específica, lo que permite una delimitación precisa del tumor y de sus subregiones internas (Zikic *et al.* , 2014; Pereira *et al.* , 2016; Abd-Ellah *et al.* , 2019).

A diferencia de las imágenes monocanal, cada vóxel en este dataset no está definido por un único valor de intensidad, sino por un vector de intensidades asociado a distintas modalidades de MRI. En concreto, cada caso está compuesto por cuatro modalidades de adquisición: T1, T1 con contraste (T1ce), T2 y FLAIR. Estas modalidades aportan información complementaria sobre el tejido cerebral, permitiendo capturar tanto la anatomía normal como las alteraciones tumorales con mayor robustez que el uso de una sola secuencia (Bauer *et al.* , 2013; Menze *et al.* , 2015; Havaei *et al.* , 2017).

Desde el punto de vista computacional, esta información se representa como un tensor de cuatro dimensiones, donde las tres primeras dimensiones corresponden a las coordenadas espaciales del volumen (X,Y,Z) y la cuarta dimensión representa los distintos canales o modalidades de imagen. De este modo, para una posición espacial concreta (x,y,z), el vóxel asociado contiene un vector de la forma:

$$[v_{T1}, v_{T1ce}, v_{T2}, v_{FLAIR}]$$

Esta representación multicanal permite a los modelos de *Deep Learning* aprender relaciones complejas entre modalidades, explotando simultáneamente información estructural, funcional y patológica del tejido cerebral (Havaei *et al.* , 2017; Isensee *et al.*, 2021).

La relevancia clínica de cada modalidad ha sido ampliamente documentada en la literatura. Mientras que la secuencia T1 proporciona una visión anatómica general del cerebro, la modalidad T1ce permite identificar regiones con captación de contraste, habitualmente asociadas a tejido tumoral activo. Por su parte, T2 resalta áreas con mayor contenido hídrico, y FLAIR suprime la señal del líquido cefalorraquídeo, facilitando la identificación del edema peritumoral y de regiones infiltrativas del tumor (Bauer *et al.* , 2013; Menze *et al.* , 2015; Abd-Ellah *et al.* , 2019).

Las imágenes incluyen anotaciones manuales realizadas por expertos, que definen las regiones tumorales de interés y sirven como referencia (*ground truth*) para el entrenamiento y la evaluación del modelo de segmentación. Estas anotaciones están alineadas espacialmente con los volúmenes de imagen, permitiendo una comparación vóxel a vóxel entre las segmentaciones predichas y las reales.

## 3.2. Preparación de los datos

El dataset original proporciona un conjunto de imágenes de entrenamiento (**imagesTr**) con sus correspondientes anotaciones (**labelsTr**), así como un conjunto de test (**imagesTs**) sin segmentaciones de referencia, lo que impide la evaluación cuantitativa directa del rendimiento del modelo sobre dicho conjunto. Por este motivo, se decidió trabajar exclusivamente con imágenes que disponen de *ground truth*, descartando el uso del conjunto imagesTs.

A partir del conjunto completo de imágenes etiquetadas, se implementó una estrategia de muestreo aleatorio mediante un script en Python, seleccionando un subconjunto de 250 pacientes para su uso en el entrenamiento del modelo. Esta selección aleatoria permite reducir posibles sesgos en la distribución de los casos y garantiza una muestra representativa de la variabilidad presente en el dataset original.

La evaluación del modelo se realizó posteriormente siguiendo el esquema de validación cruzada integrado en nnU-Net, sin necesidad de definir manualmente un conjunto de validación independiente.

```
seleccionados = random.sample(
    pacientes_disponibles,
    min(CANTIDAD_MUESTRA, len(pacientes_disponibles))
)
```

## 3.3. Separación de canales y reorganización del dataset

Con el objetivo de adaptar el conjunto de datos al formato requerido por nnU-Net, se desarrolló un script en Python que automatiza la carga y el procesamiento de los volúmenes MRI multicanal. Cada archivo NIfTI original, que contiene internamente información de múltiples modalidades, fue cargado y descompuesto en cuatro volúmenes tridimensionales independientes, correspondientes a los distintos canales del volumen original.

Cada uno de estos volúmenes 3D fue almacenado siguiendo estrictamente la convención de nombres obligatoria de nnU-Net (ID_0000.nii.gz, ID_0001.nii.gz, ID_0002.nii.gz e ID_0003.nii.gz), lo que permite al framework identificar automáticamente las modalidades de entrada asociadas a cada paciente. Durante este proceso se conservaron la matriz de transformación espacial (*affine*) y los metadatos originales del archivo NIfTI, garantizando la coherencia geométrica entre las imágenes y las máscaras de segmentación.

De forma paralela, las máscaras de segmentación asociadas a cada paciente fueron copiadas a la estructura de directorios correspondiente (labelsTr), verificando en cada caso la existencia de la anotación para asegurar la integridad del conjunto de datos empleado.

## 3.4. Definición del dataset para nnU-Net

Una vez reorganizados los datos, se generó el archivo de configuración dataset.json, requerido por nnU-Net para describir formalmente el conjunto de datos. En este archivo se definieron explícitamente los canales de entrada, asignando cada índice a su modalidad correspondiente (T1, T1ce, T2 y FLAIR), así como las etiquetas de segmentación empleadas en el estudio.

El archivo dataset.json incluye además el número de casos utilizados para el entrenamiento, el formato de los archivos de entrada y una descripción del dataset experimental, permitiendo a nnU-Net inferir automáticamente la dimensionalidad de los datos y configurar su pipeline de preprocesamiento y entrenamiento sin intervención manual adicional.

## 3.5. Preprocesamiento automático con nnU-Net

El preprocesamiento de los datos se llevó a cabo utilizando el pipeline estándar de nnU-Net (nnUNetv2_plan_and_preprocess) bajo la configuración 3d_fullres. Durante esta etapa, el framework extrae automáticamente propiedades relevantes del conjunto de datos, como el espaciado de vóxels, las dimensiones espaciales de los volúmenes y la distribución de intensidades.

Asimismo, nnU-Net realiza de forma automática el diseño de la arquitectura de la red neuronal en función de estas propiedades, ajustando parámetros como el tamaño de los parches de entrada, la profundidad de la red y las estrategias de normalización. El

preprocesamiento incluye el recorte de regiones relevantes, la normalización de intensidades mediante z-score y el almacenamiento eficiente de los datos en formatos optimizados para el entrenamiento.

Previamente al inicio del entrenamiento, se ejecutó una verificación de integridad del dataset para asegurar la correspondencia correcta entre imágenes y etiquetas, garantizando la consistencia del conjunto de datos empleado.

```
nnUNetv2_plan_and_preprocess -d 501 -c 3d_fullres --
verify_dataset_integrity
```

## 3.6. Entrenamiento del modelo

El entrenamiento del modelo se realizó utilizando la implementación nnU-Net v2, empleando una arquitectura U-Net tridimensional generada automáticamente por el framework en función de las características del conjunto de datos. Se utilizó la configuración 3d_fullres, que permite entrenar el modelo utilizando la resolución espacial completa de los volúmenes con el objetivo de maximizar la precisión en la segmentación de las regiones tumorales.

El proceso de entrenamiento se ejecutó mediante el comando nnUNetv2_train, especificando el dataset ID 501 y entrenando el primer pliegue (fold 0) del esquema de validación cruzada de cinco pliegues (5-fold cross-validation) implementado por nnU-Net. En este esquema, aproximadamente el 80 % de los datos se emplea para el entrenamiento y el 20 % restante para la validación interna del modelo.

Debido al aumento significativo del tamaño del conjunto de entrenamiento respecto a pruebas preliminares, se incrementó el número de épocas de entrenamiento con el fin de asegurar una correcta convergencia del modelo y permitir la estabilización de las métricas de segmentación. Todo el proceso de entrenamiento se llevó a cabo en un entorno con disponibilidad de GPU, concretamente utilizando **Google Colab** con aceleración por hardware, condición indispensable para el entrenamiento eficiente de redes neuronales profundas sobre volúmenes tridimensionales multicanal.

```
!nnUNetv2_train 501 3d_fullres 0
```

## 3.7. Verificación del aprendizaje del modelo

Con el objetivo de comprobar que el modelo había aprendido patrones relevantes durante el entrenamiento, se analizó la gráfica de progreso generada automáticamente por nnU-Net. Esta gráfica recoge la evolución de la función de pérdida (loss) y de la métrica de segmentación (*Dice score*) tanto para el conjunto de entrenamiento como para el conjunto de validación a lo largo de las épocas.

La inspección visual de estas curvas permite evaluar la dinámica de aprendizaje del modelo, identificar el punto de convergencia y detectar posibles fenómenos de sobreajuste. En particular, una disminución progresiva de la pérdida de entrenamiento

junto con la estabilización o mejora del *Dice score* de validación constituye una evidencia empírica de que el modelo está capturando correctamente las características relevantes del problema de segmentación (Bakas *et al.* , 2017).

Para facilitar este análisis, la gráfica de progreso fue cargada y visualizada directamente desde el directorio de resultados generado por nnU-Net.

## 3.8. Preparación del conjunto de prueba independiente

Con el objetivo de evaluar el rendimiento del modelo sobre datos no utilizados durante el entrenamiento, se preparó un conjunto de prueba independiente a partir del dataset original. Para ello, se identificaron automáticamente los pacientes empleados en el entrenamiento a partir de la estructura de directorios generada por nnU-Net, excluyendo dichos casos del conjunto de prueba.

```
ids_entrenados = set()
for f in archivos_train:
    id_clean = "_".join(f.split('_')[:-1])
    ids_entrenados.add(id_clean)
```

A continuación, se seleccionaron los pacientes restantes disponibles en el dataset original que no habían sido utilizados durante el entrenamiento. Para cada uno de estos sujetos, las imágenes de resonancia magnética fueron procesadas de forma análoga al conjunto de entrenamiento, cargando los volúmenes NIfTI multicanal y separando explícitamente cada modalidad en volúmenes tridimensionales independientes siguiendo la convención de nombres requerida por nnU-Net.

```
if id_paciente not in ids_entrenados:
    pacientes_para_test.append(p)
```

Las imágenes procesadas se organizaron en un directorio específico destinado a servir como entrada del modelo durante la fase de inferencia, mientras que las máscaras de segmentación correspondientes (*ground truth*) se almacenaron de forma separada para su uso posterior en la evaluación cuantitativa del rendimiento del modelo mediante métricas de segmentación.

```
volumen = data[:, :, :, canal]
new_img = nib.Nifti1Image(volumen, affine, header)
nib.save(new_img, os.path.join(path_test_input, nombre_final))
```

## 3.9. Inferencia y generación de segmentaciones automáticas

Una vez entrenado el modelo y seleccionado el estado óptimo correspondiente al mejor rendimiento durante la validación, se procedió a la generación de segmentaciones automáticas sobre el conjunto de prueba independiente. La inferencia se realizó utilizando el comando nnUNetv2_predict, empleando la misma configuración 3d_fullres y el mismo pliegue (*fold* 0) utilizados durante el entrenamiento.

```
!nnUNetv2_predict -i input_folder -o output_folder \
-d 501 -c 3d_fullres -f 0 -chk checkpoint_best.pth
```

Con el fin de maximizar la capacidad de generalización del modelo, se especificó explícitamente el uso del *checkpoint* correspondiente a la época de mejor rendimiento (checkpoint_best), descartando el estado final del entrenamiento. Las imágenes de entrada, previamente organizadas siguiendo la estructura requerida por nnU-Net, fueron procesadas automáticamente para generar las máscaras de segmentación predichas.

Durante este proceso se registró el tiempo total de ejecución de la inferencia, permitiendo obtener una estimación del coste computacional asociado a la aplicación del modelo entrenado. Las máscaras de segmentación resultantes se almacenaron en un directorio específico para su posterior análisis cualitativo y cuantitativo en el apartado de Resultados.

## 3.10. Métricas de evaluación del modelo

### 3.10.1. Coeficiente Dice

Con el objetivo de evaluar cuantitativamente el grado de solapamiento entre las segmentaciones generadas por el modelo y las segmentaciones manuales de referencia, se empleó el coeficiente Dice. Esta métrica es ampliamente utilizada en tareas de segmentación biomédica, ya que proporciona una medida robusta de la concordancia espacial entre dos regiones segmentadas (Taha & Hanbury, 2015).

El coeficiente Dice se calculó a nivel vóxel comparando, para cada paciente del conjunto de prueba independiente, la máscara de segmentación predicha por el modelo con su correspondiente *ground truth*. Con el fin de evaluar el volumen tumoral completo, las máscaras fueron binarizadas considerando como región tumoral todos los vóxeles con valor mayor que cero, integrando de este modo las distintas subregiones tumorales en una única clase.

El cálculo del coeficiente Dice se realizó de forma automática mediante un procedimiento computacional que recorrió secuencialmente todas las predicciones disponibles, almacenando el valor de la métrica para cada paciente.

### 3.10.2. Evaluación volumétrica del tumor

Además del solapamiento espacial, se evaluó la capacidad del modelo para estimar correctamente el volumen tumoral, una magnitud de especial relevancia clínica en el seguimiento de la enfermedad (Bakas *et al.* , 2017). Para ello, se comparó el volumen tumoral total obtenido a partir de las segmentaciones manuales de referencia con el volumen estimado por el modelo.

En ambos casos, el volumen se calculó como el número total de vóxeles clasificados como tumor tras la binarización de las máscaras de segmentación. Esta aproximación permite analizar la consistencia global de las predicciones del modelo en términos de

tamaño tumoral, independientemente de la forma específica de las regiones segmentadas.

#### 3.10.3. Distancia de Hausdorff al 95 % (HD95)

Con el objetivo de evaluar la precisión en la delimitación de los bordes tumorales, se empleó la distancia de Hausdorff al 95 % (HD95). Esta métrica cuantifica la discrepancia espacial entre los contornos de la segmentación automática y la segmentación manual de referencia, descartando el 5 % de los valores extremos para reducir la influencia de outliers (Taha & Hanbury, 2015).

El cálculo del HD95 se realizó sobre las segmentaciones binarizadas, expresando la distancia en milímetros mediante el uso del espaciado real de los vóxeles extraído de los metadatos de cada volumen NIfTI, lo que garantiza una interpretación física directa del error en la localización de los bordes tumorales.

## 3.11. Criterio de selección de casos

Con el objetivo de realizar una evaluación cualitativa representativa y reproducible del rendimiento del modelo, se definió un criterio objetivo para la selección de los casos analizados visualmente. Concretamente, los pacientes del conjunto de prueba independiente fueron ordenados automáticamente en función del coeficiente Dice obtenido durante la evaluación cuantitativa.

A partir de este ordenamiento, se seleccionaron los cinco pacientes con mayor Dice score (top-5), correspondientes a los casos con mayor concordancia entre la segmentación automática y la segmentación manual de referencia. Este procedimiento evita sesgos derivados de una selección manual de ejemplos y permite ilustrar el comportamiento del modelo en escenarios de alto rendimiento.

# 4. Resultados

## 4.1. Evidencia de aprendizaje del modelo

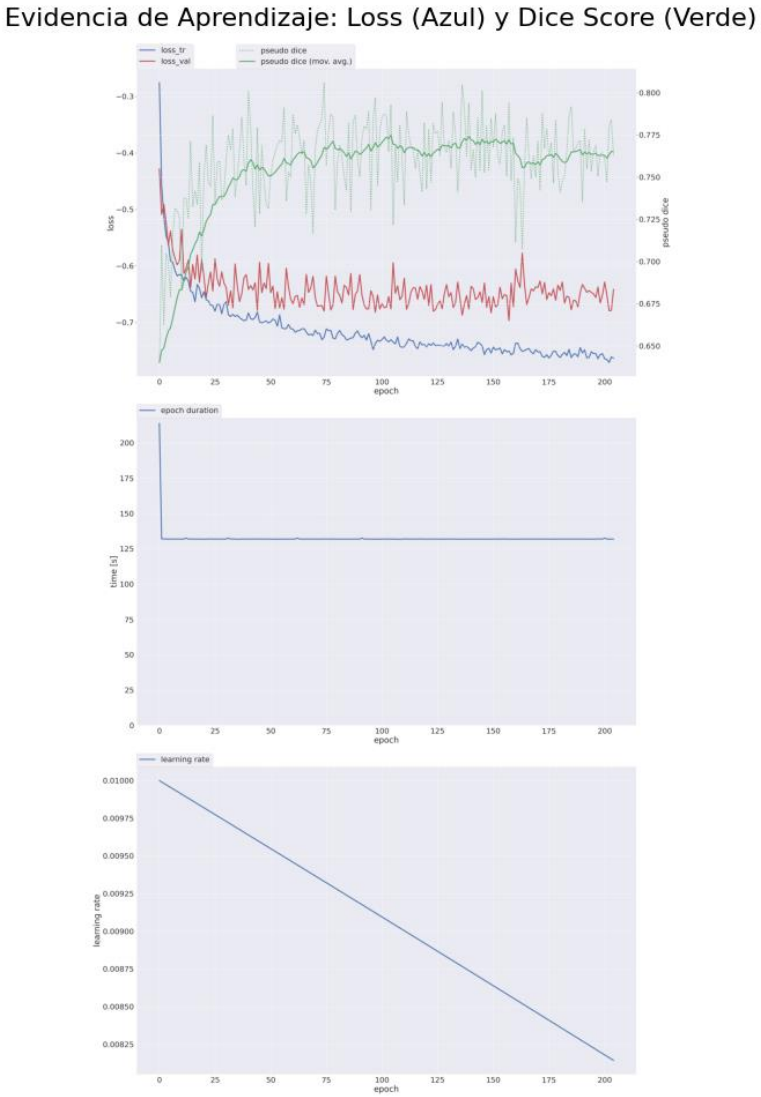



***Figura 1. Evolución del proceso de entrenamiento del modelo nnU-Net 3D (3d_fullres).*** *En la gráfica superior se muestran la pérdida de entrenamiento y validación (izquierda) junto con la métrica* pseudo Dice *(derecha). La gráfica central representa la duración de cada época, mientras que la gráfica inferior muestra la evolución de la tasa de aprendizaje a lo largo de las 200 épocas de entrenamiento.*

Con el objetivo de verificar que el modelo entrenado ha aprendido patrones relevantes para la tarea de segmentación tumoral, se analizó la evolución de las métricas de entrenamiento y validación generadas automáticamente por el *framework* nnU-Net durante el proceso de entrenamiento. La **Figura 1** muestra la gráfica de progreso correspondiente al entrenamiento del modelo U-Net 3D bajo la configuración *3d_fullres*.

En la gráfica superior se representa la evolución de la función de pérdida (*loss*) tanto para el conjunto de entrenamiento como para el conjunto de validación, junto con la métrica de segmentación *pseudo Dice*. Durante las primeras épocas se observa una disminución progresiva de la pérdida de entrenamiento acompañada de un incremento

sostenido del *pseudo Dice*, lo que indica que el modelo está aprendiendo de forma efectiva a partir de los datos de entrada.

A medida que avanza el entrenamiento, la pérdida de validación tiende a estabilizarse, mientras que la métrica *pseudo Dice* alcanza un máximo claramente definido. Este comportamiento es característico de un proceso de convergencia adecuado y permite identificar el punto óptimo de entrenamiento, previo a la aparición de fenómenos de sobreajuste (*overfitting*). A partir de dicho punto, las mejoras adicionales en la pérdida de entrenamiento no se traducen en un incremento del rendimiento sobre los datos de validación.

La gráfica central muestra la duración de cada época de entrenamiento, que se mantiene prácticamente constante a lo largo de todo el proceso. Esta estabilidad descarta la presencia de anomalías computacionales o cuellos de botella en la carga de datos, indicando que las variaciones observadas en las métricas responden exclusivamente a la dinámica de aprendizaje del modelo.

Por último, la gráfica inferior representa la evolución de la tasa de aprendizaje (*learning rate*), que presenta una reducción gradual a lo largo de las épocas. Este comportamiento contribuye a la estabilización del entrenamiento en fases avanzadas, permitiendo ajustes más finos de los pesos de la red neuronal conforme el modelo se aproxima a su punto de convergencia.

## 4.2. Resultados cuantitativos del modelo

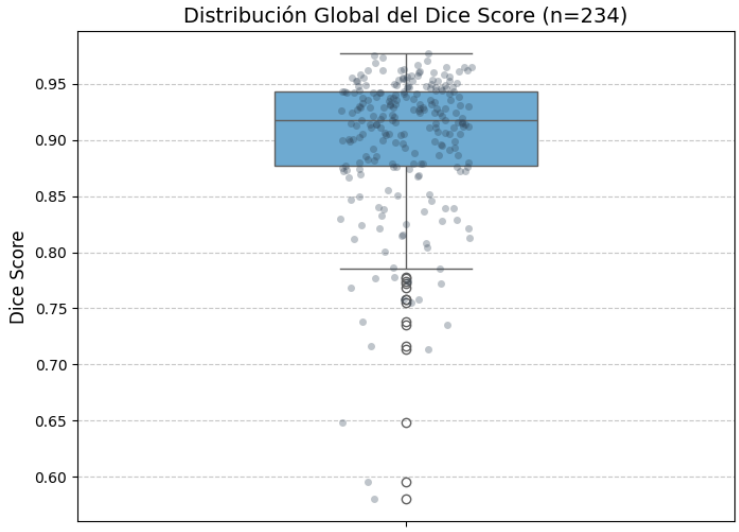



***Figura 2. Distribución global del Dice score obtenido para el conjunto de prueba independiente (n = 234).*** *El diagrama de caja representa la mediana, el rango intercuartílico y los valores extremos, mientras que los puntos individuales muestran el Dice score correspondiente a cada paciente evaluado.*

La **Figura 2** muestra la distribución global del *Dice score* obtenido para todos los pacientes evaluados. El modelo alcanza una **media de *Dice score* de 0.8997** y una **mediana de 0.9174**, lo que indica un alto grado de solapamiento entre las segmentaciones predichas y las anotaciones manuales en la mayoría de los casos. La

**desviación estándar**, igual a **0.0636**, refleja una variabilidad moderada en el rendimiento del modelo entre distintos pacientes.

La distribución presenta un rango intercuartílico relativamente estrecho, concentrándose la mayor parte de los valores de Dice por encima de 0.88, lo que sugiere un comportamiento consistente del modelo en la segmentación del volumen tumoral completo. No obstante, se observan valores atípicos con *Dice score*s inferiores, correspondientes a un reducido número de casos en los que el modelo presenta mayores dificultades de segmentación.

## 4.3. Análisis de la estimación del volumen tumoral

La **Figura 3** muestra la correlación entre el volumen real del tumor y el volumen estimado por el modelo para todos los pacientes del conjunto de prueba. Cada punto representa un caso individual, mientras que la recta de regresión ilustra la relación lineal entre ambas magnitudes.

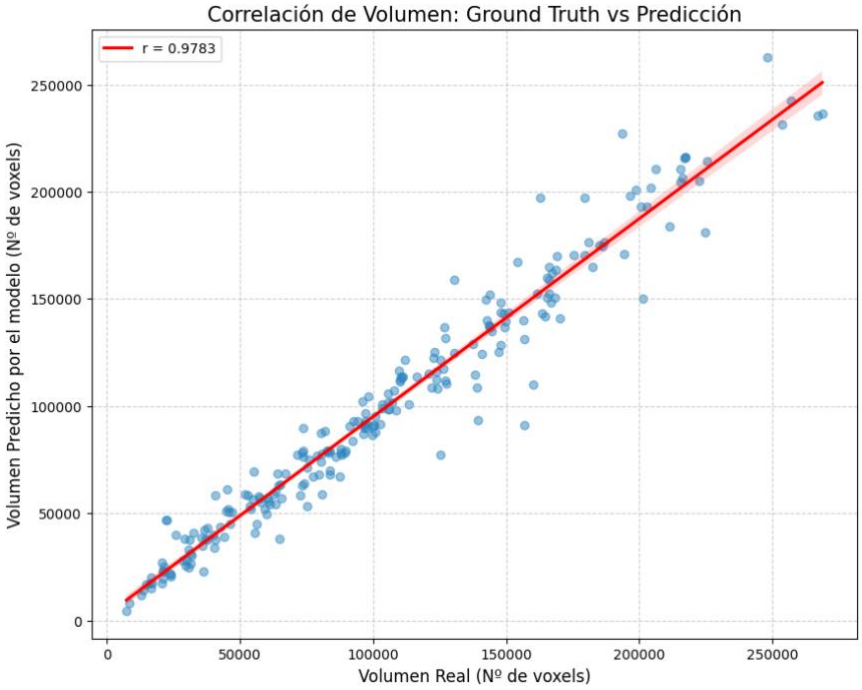



***Figura 3. Correlación entre el volumen tumoral real y el volumen estimado por el modelo para el conjunto de prueba independiente.*** *Cada punto representa un paciente, mientras que la recta roja corresponde a la regresión lineal. Se indica el coeficiente de correlación de Pearson (r).*

El análisis revela una **correlación positiva muy elevada** entre el volumen real y el volumen predicho, con un **coeficiente de correlación de Pearson r = 0.9783**, **estadísticamente significativo** ($p = 1.47 \times 10^{-160}$). Este resultado indica que el modelo no solo segmenta correctamente la localización del tumor, sino que también preserva de forma consistente la proporcionalidad del volumen tumoral en un amplio rango de tamaños. La cercanía de los puntos a la recta de regresión sugiere que los errores de sobresegmentación y subsegmentación tienden a compensarse globalmente, dando lugar a estimaciones volumétricas robustas incluso en tumores de gran tamaño. No obstante, se observa una ligera dispersión en los casos con volúmenes más elevados, lo que podría estar relacionado con la mayor complejidad morfológica de estos tumores.

## 4.4. Evaluación del error en los bordes mediante la distancia de Hausdorff (HD95)

El análisis se aplicó a todos los casos del conjunto de prueba con segmentación de referencia, obteniéndose un valor medio de HD95 de 5.97 mm. Este resultado indica que, en promedio, el 95 % del contorno predicho se encuentra a menos de 6 mm del contorno real, lo que refleja una delimitación espacial precisa de los márgenes tumorales en la mayoría de los casos, teniendo en cuenta la naturaleza difusa de los bordes tumorales en MRI.

## 4.5. Resultados cualitativos de la segmentación

### 4.5.1. Comparación visual FLAIR–GT–Predicción

Con el fin de evaluar la coherencia anatómica de las segmentaciones generadas por el modelo, se realizó una comparación visual directa entre la imagen original, la segmentación manual de referencia y la predicción automática. Para esta comparación se utilizó la secuencia FLAIR, dado que es especialmente informativa para la visualización de tumores cerebrales y regiones de edema peritumoral.

Para cada paciente seleccionado, se representó un corte axial del volumen, escogido automáticamente como aquel que contiene el mayor volumen tumoral según la segmentación de referencia. Esta estrategia garantiza que la comparación visual se centre en la región anatómica de mayor relevancia clínica y evita una selección arbitraria de cortes.

En la **Figura 4** se muestran, de izquierda a derecha, las imágenes FLAIR original en escala de grises, la segmentación manual de referencia (*ground truth*) superpuesta sobre la imagen anatómica y la segmentación generada por el modelo nnU-Net. Esta disposición permite una comparación directa entre la delimitación clínica del tumor y la predicción automática, facilitando la evaluación visual del grado de concordancia entre ambas.

En los casos analizados se observa una elevada correspondencia en la localización del tumor y en la delimitación de sus regiones principales. El modelo reproduce de forma consistente la forma global del tumor y su extensión espacial, mostrando una buena alineación con la segmentación manual. Las discrepancias visuales se concentran principalmente en los bordes del tumor, donde la transición entre tejido sano y patológico es más difusa incluso para observadores expertos.

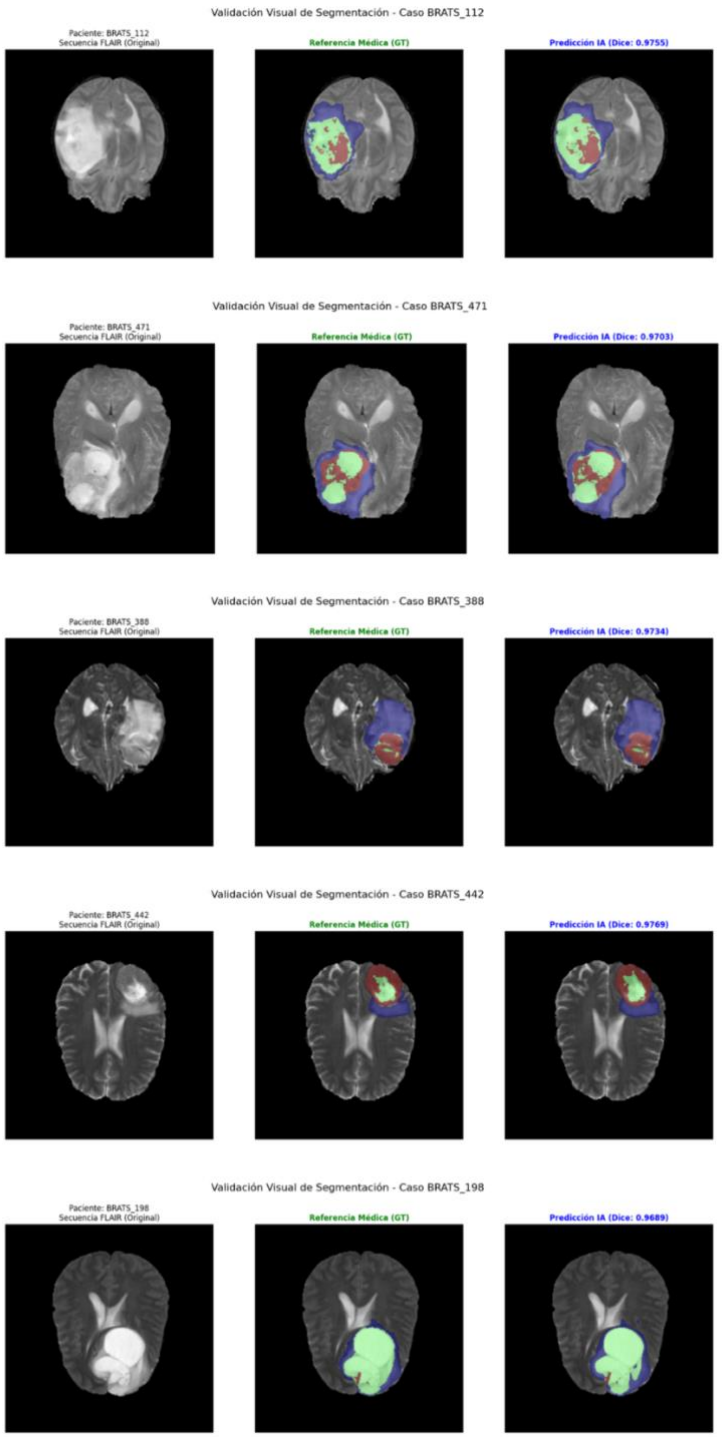



***Figura 4. Comparación visual entre imagen FLAIR original, segmentación manual de referencia (ground truth) y segmentación generada por el modelo nnU-Net*** *para las cinco mejores predicciones del conjunto de prueba con mayor Dice score. Para cada caso se muestra un corte axial seleccionado automáticamente como aquel con mayor volumen tumoral.*

### 4.5.2. Análisis visual de errores de segmentación

Con el fin de identificar de forma explícita las limitaciones del modelo y comprender la naturaleza de los errores cometidos durante la segmentación automática, se realizó un análisis visual detallado de los errores de segmentación tanto en casos de alto rendimiento como en los casos con peor desempeño cuantitativo.

Para este análisis se generaron mapas de error superpuestos sobre la imagen anatómica original en secuencia FLAIR. En dichos mapas se distinguieron tres regiones fundamentales:

(i) **aciertos o verdaderos positivos**, correspondientes a vóxeles correctamente segmentados por el modelo;
(ii) **omisiones o falsos negativos**, donde el modelo no detecta regiones tumorales presentes en la segmentación manual; y
(iii) **sobresegmentaciones o falsos positivos**, donde el modelo clasifica erróneamente tejido sano como tumoral.

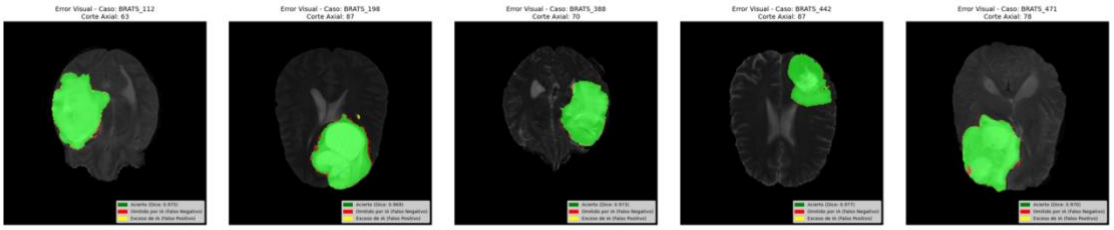


***Figura 5. Análisis visual de errores de segmentación para un caso de alto rendimiento.*** *Las regiones correctamente segmentadas se muestran en verde, mientras que las omisiones (falsos negativos) y las sobresegmentaciones (falsos positivos) se indican en rojo y amarillo, respectivamente, superpuestas sobre la imagen FLAIR original.*

En primer lugar, se analizaron los cinco pacientes con mayor *Dice score* **(Figura 5)**, con el objetivo de evaluar cómo se distribuyen los errores incluso en escenarios de alto rendimiento. En estos casos, la mayor parte del volumen tumoral aparece correctamente segmentado, observándose que los errores se concentran principalmente en los bordes del tumor, donde la transición entre tejido sano y patológico es progresiva y presenta una elevada ambigüedad. Este patrón indica que el modelo aprende de forma robusta las regiones centrales del tumor, mientras que las discrepancias se limitan a zonas periféricas de difícil delimitación incluso para expertos humanos.

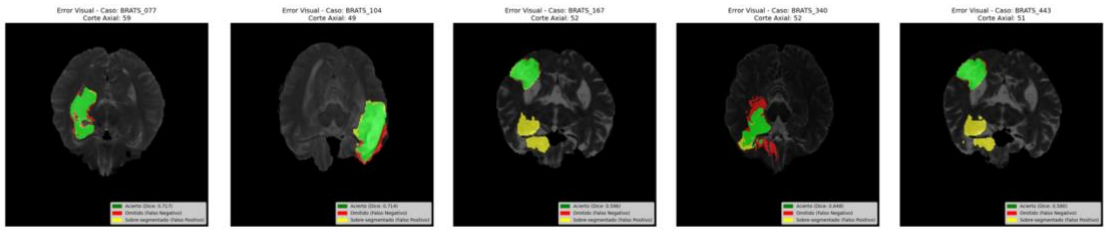


***Figura 6. Mapas de error de segmentación correspondientes a los cinco pacientes con menor Dice score del conjunto de prueba.*** *Las regiones correctamente segmentadas se muestran en verde, mientras que las omisiones (falsos negativos) y las sobresegmentaciones (falsos positivos) se indican en rojo y amarillo, respectivamente, superpuestas sobre imágenes FLAIR originales.*

La **Figura 6** muestra los mapas de error de segmentación correspondientes a los cinco pacientes con menor *Dice score* del conjunto de prueba. En estas visualizaciones, las regiones correctamente segmentadas por el modelo se representan en verde, mientras que las omisiones (falsos negativos) aparecen en rojo y las sobresegmentaciones (falsos positivos) en amarillo, superpuestas sobre la imagen anatómica original en secuencia FLAIR.

En todos los casos analizados se observa que, a pesar de un rendimiento cuantitativo inferior, el modelo mantiene una localización general correcta del tumor, como evidencia

la presencia de regiones verdes que coinciden con el núcleo tumoral principal. No obstante, los errores adquieren una mayor extensión y relevancia clínica en comparación con los casos de alto rendimiento.

Un patrón recurrente en estos mapas de error es la presencia de omisiones parciales en regiones periféricas del tumor, especialmente en zonas con bajo contraste respecto al tejido sano circundante. Estas áreas, marcadas en rojo, indican que el modelo tiende a infraestimar la extensión tumoral en regiones donde la transición entre tejido patológico y tejido sano es gradual o difusa.

Asimismo, se identifican regiones de sobresegmentación (amarillo), localizadas principalmente en estructuras adyacentes al tumor o en áreas con intensidades similares al tejido tumoral en la secuencia FLAIR. Este comportamiento sugiere que el modelo puede confundir determinadas regiones anatómicas con edema o tejido tumoral, especialmente en casos con morfologías complejas o tumores fragmentados.

Cabe destacar que los errores no se distribuyen de forma homogénea en todo el volumen, sino que se concentran en zonas anatómicamente complejas, como regiones cercanas a los ventrículos o áreas profundas del encéfalo. Este patrón es consistente con la literatura previa y pone de manifiesto las limitaciones inherentes a la segmentación automática en escenarios clínicamente ambiguos.

## 5. Discusión

Los resultados obtenidos en este estudio evidencian que el modelo nnU-Net entrenado es capaz de aprender representaciones relevantes para la segmentación automática de tumores cerebrales en imágenes de resonancia magnética multimodal.

La **evolución conjunta de la función de pérdida** y de la métrica pseudo Dice durante el entrenamiento muestra un comportamiento característico de convergencia estable, con una disminución progresiva de la pérdida y una estabilización del rendimiento en validación. Este patrón sugiere que el modelo logra un equilibrio adecuado entre capacidad de aprendizaje y generalización, sin manifestar signos claros de sobreajuste, lo que concuerda con el diseño autoajustable de nnU-Net y con observaciones previas en la literatura sobre su robustez frente a variaciones en el dataset y en los hiperparámetros de entrenamiento (Isensee *et al.* , 2021; Bakas *et al.* , 2017).

El **análisis cuantitativo** mediante el coeficiente Dice revela un alto grado de solapamiento entre las segmentaciones generadas por el modelo y las anotaciones manuales de referencia. La media de Dice obtenida, cercana a 0.9, junto con una mediana aún más elevada, indica que el modelo segmenta correctamente el volumen tumoral completo en la mayoría de los casos evaluados. Este nivel de rendimiento es comparable al reportado por otros enfoques basados en redes neuronales convolucionales profundas aplicadas a datasets similares, lo que refuerza la validez del enfoque adoptado (Havaei *et al.* , 2017; Pereira *et al.* , 2016). La presencia de una desviación estándar moderada sugiere, no obstante, que el rendimiento no es

completamente homogéneo entre pacientes, reflejando la elevada heterogeneidad morfológica y de intensidad característica de los gliomas en imágenes MRI (Menze *et al.* , 2015).

La **distribución del Dice score** muestra que la mayor parte de los casos se concentra en valores elevados, mientras que los valores atípicos con bajo rendimiento corresponden a un número reducido de pacientes. Este comportamiento es consistente con observaciones previas que indican que los casos más problemáticos suelen estar asociados a tumores con bordes especialmente difusos, regiones infiltrativas extensas o bajo contraste entre tejido tumoral y tejido sano, incluso en secuencias como FLAIR (Abd-Ellah *et al.* , 2019; Bauer *et al.* , 2013). Estos resultados ponen de manifiesto que, aunque el modelo presenta un rendimiento global elevado, existen escenarios clínicamente complejos en los que la segmentación automática sigue siendo un desafío.

El **análisis de la estimación del volumen tumoral** complementa la evaluación basada en métricas de solapamiento, aportando una perspectiva clínicamente relevante. La elevada correlación observada entre el volumen real y el volumen estimado por el modelo indica que nnU-Net no solo segmenta correctamente la localización del tumor, sino que también preserva de forma consistente la proporcionalidad del volumen tumoral en un amplio rango de tamaños. Este resultado es especialmente relevante en contextos clínicos donde la estimación precisa del volumen tumoral es fundamental para la planificación terapéutica y el seguimiento de la evolución de la enfermedad (Menze *et al.* , 2015). La ligera dispersión observada en tumores de gran tamaño puede atribuirse a la mayor complejidad morfológica de estos casos, donde pequeñas discrepancias en los bordes pueden traducirse en diferencias volumétricas apreciables.

La **evaluación del error en los bordes** mediante la distancia de Hausdorff al 95 % (HD95) proporciona información adicional sobre la precisión geométrica de las segmentaciones generadas. El valor medio obtenido, inferior a 6 mm, indica que la mayor parte de los contornos predichos se sitúa muy próxima a la segmentación manual de referencia, lo que sugiere una delimitación precisa de los márgenes tumorales. Este resultado es coherente con estudios previos que destacan la importancia de complementar métricas de solapamiento con métricas basadas en distancias para evaluar correctamente la calidad de las segmentaciones, especialmente en regiones donde pequeños errores pueden tener implicaciones clínicas significativas (Bauer *et al.* , 2013; Zikic *et al.* , 2014).

**Los resultados cualitativos** refuerzan las conclusiones extraídas a partir de las métricas cuantitativas. La comparación visual entre las imágenes FLAIR originales, las segmentaciones manuales y las predicciones automáticas muestra una elevada concordancia en la localización y forma global del tumor en los casos con mayor Dice score. En estos escenarios, el modelo reproduce de forma consistente las regiones centrales del tumor y el edema peritumoral, lo que sugiere que ha aprendido representaciones robustas de las características más discriminativas del tejido tumoral en imágenes multimodales (Havaei *et al.* , 2017; Isensee *et al.* , 2021).

**El análisis visual de errores** revela patrones recurrentes que ayudan a comprender las limitaciones del modelo. Incluso en los **casos de alto rendimiento**, los errores tienden a

concentrarse en los bordes del tumor, donde la transición entre tejido sano y patológico es gradual y presenta una elevada ambigüedad. Este fenómeno ha sido ampliamente documentado en la literatura y refleja una dificultad inherente tanto para los modelos automáticos como para los observadores humanos, especialmente en tumores infiltrativos característicos de los gliomas de alto grado (Menze *et al.* , 2015; Abd-Ellah *et al.* , 2019).

En los casos con **peor desempeño cuantitativo**, los mapas de error muestran una mayor presencia de omisiones parciales y sobresegmentaciones localizadas, particularmente en regiones anatómicamente complejas o con intensidades similares al tejido tumoral. La tendencia a infraestimar la extensión tumoral en áreas periféricas de bajo contraste sugiere que el modelo prioriza regiones con señales más claras, mientras que las sobresegmentaciones en estructuras adyacentes pueden deberse a la similitud de intensidades en determinadas secuencias MRI, especialmente en FLAIR (Bauer *et al.* , 2013; Pereira *et al.* , 2016).

En conjunto, los resultados obtenidos indican que el uso de nnU-Net para la segmentación automática de tumores cerebrales en MRI multimodal permite alcanzar un rendimiento elevado y consistente, tanto desde el punto de vista cuantitativo como cualitativo. No obstante, las limitaciones observadas en la delimitación de bordes y en casos morfológicamente complejos ponen de manifiesto la necesidad de continuar explorando estrategias que mejoren la sensibilidad del modelo en regiones ambiguas, como la incorporación de información contextual adicional o el refinamiento de las estrategias de postprocesado (Isensee *et al.* , 2021; Ottom *et al.* , 2022).

## 6. Conclusiones

1. En este trabajo se ha desarrollado y evaluado un modelo de segmentación automática de tumores cerebrales basado en **nnU-Net**, utilizando imágenes de resonancia magnética multimodal procedentes del dataset público BraTS. El enfoque adoptado permite abordar la segmentación vóxel a vóxel de volúmenes tridimensionales complejos de forma completamente automática.
2. El **análisis del proceso de entrenamiento** ha mostrado un comportamiento de convergencia estable, caracterizado por una disminución progresiva de la función de pérdida y una estabilización de la métrica Dice en el conjunto de validación. Este patrón indica que el modelo ha aprendido representaciones relevantes del problema sin evidencias claras de sobreajuste.
3. **La evaluación cuantitativa del rendimiento** del modelo ha puesto de manifiesto un alto grado de solapamiento entre las segmentaciones predichas y las segmentaciones manuales de referencia, con valores elevados del coeficiente Dice en la mayoría de los casos evaluados. Estos resultados confirman la capacidad del modelo para segmentar correctamente el volumen tumoral completo de forma consistente.
4. El **análisis volumétrico** ha revelado una correlación muy elevada entre el volumen tumoral real y el volumen estimado por el modelo, lo que indica que la segmentación automática preserva adecuadamente la proporcionalidad del

tamaño tumoral en un amplio rango de volúmenes. Este aspecto resulta especialmente relevante desde el punto de vista clínico para la planificación terapéutica y el seguimiento de la enfermedad.

5. **La evaluación de la precisión en los bordes** mediante la distancia de Hausdorff al 95 % ha mostrado errores espaciales reducidos, lo que sugiere una delimitación precisa de los márgenes tumorales en la mayoría de los casos, aunque persisten discrepancias localizadas en regiones anatómicamente complejas.
6. **Los resultados cualitativos y el análisis visual** de errores han permitido identificar patrones recurrentes en las limitaciones del modelo, principalmente asociados a bordes tumorales difusos, regiones infiltrativas y áreas con bajo contraste entre tejido sano y patológico. Estas limitaciones reflejan desafíos inherentes a la segmentación automática de tumores cerebrales en MRI.
7. En conjunto, los resultados obtenidos demuestran que nnU-Net constituye una herramienta robusta y eficaz para la segmentación automática de tumores cerebrales en imágenes MRI multimodales. No obstante, futuras líneas de trabajo podrían centrarse en mejorar la precisión en la delimitación de bordes y en casos de alta complejidad morfológica, así como en la evaluación del modelo en escenarios clínicos más heterogéneos.

# 7. BIBLIOGRAFÍA


Abd-Ellah, M. K., Awad, A. I., Khalaf, A. A. M., & Hamed, H. F. A. (2019). A review on brain tumor diagnosis from MRI images: Practical implications, key achievements, and lessons learned. *Magnetic Resonance Imaging, 61*, 300–318. https://doi.org/10.1016/j.mri.2019.01.028

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., Freymann, J. B., Farahani, K., & Davatzikos, C. (2017). Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data, 4*, 170117. https://doi.org/10.1038/sdata.2017.117

Bauer, S., Wiest, R., Nolte, L.-P., & Reyes, M. (2013). A survey of MRI-based medical image analysis for brain tumor studies. *Physics in Medicine & Biology, 58*(13), R97–R129. https://doi.org/10.1088/0031-9155/58/13/R97

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., & Larochelle, H. (2017). Brain tumor segmentation with deep neural networks. *Medical Image Analysis, 35*, 18–31. https://doi.org/10.1016/j.media.2016.05.004

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods, 18*, 203–211. https://doi.org/10.1038/s41592-020-01008-z

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., & Reyes, M. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging, 34*(10), 1993–2024. https://doi.org/10.1109/TMI.2014.2377694

Ottom, M. A., Abdul Rahman, H., & Dinov, I. D. (2022). Znet: Deep learning approach for 2D MRI brain tumor segmentation. *IEEE Journal of Translational Engineering in Health and Medicine, 10*, 1–14. https://doi.org/10.1109/JTEHM.2022.3176737

Pereira, S., Pinto, A., Alves, V., & Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Transactions on Medical Imaging, 35*(5), 1240–1251. https://doi.org/10.1109/TMI.2016.2538465

Taha, A. A., & Hanbury, A. (2015). . **BMC Medical Imaging, 15**, Article 29. https://doi.org/10.1186/s12880-015-0068-x

Zikic, D., Glocker, B., Konukoglu, E., Criminisi, A., Demiralp, Ç., Shotton, J., & Price, S. J. (2014). Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel MR. *Medical Image Analysis, 18*(1), 12–25. https://doi.org/10.1016/j.media.2013.10.002