Microservices architecture for automated processing and advanced three-dimensional segmentation of glioblastoma

Alexander Mulet de los Reyes[1,2,*], Karina Breitburd[3], Julio Jacobo Berlles[4], María Elena Buemi[4] and Cecilia Suárez[1,2]

[1] Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Física, Buenos Aires, Argentina.
[2] Instituto de Física Interdisciplinaria y Aplicada (INFINA), UBA-CONICET , Buenos Aires, Argentina.
[3] Hospital Alemán, Servicio de Neurocirugía, Buenos Aires, Argentina
[4] Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Computación. Buenos Aires, Argentina.
* Corresponding author: amuletdelosreyes@df.uba.ar

Abstract. This paper presents the development of a scalable platform based on microservices in containers for the automated analysis of glioblastoma multiparametric magnetic resonance imaging. The system includes a web interface that manages the dynamic ingestion of imaging modalities, bridging the gap between the clinical and computational environments. The workflow begins with standardized image preprocessing based on the BraTS challenge protocol. This process includes converting DICOM formats to NIfTI, reorienting the coordinate system, and performing rigid registration of the T1-Gd sequence to an anatomical reference atlas at an isotropic resolution of 1 mm^3. Subsequently, the T1, T2, and T2-FLAIR sequences are co-registered to the T1-Gd sequence, and brain tissue is extracted using algorithmically generated masks. The analytical core employs deep neural networks (SwinUNETR transformers followed by contrastive networks) through two concurrent inference pipelines. The first distinguishes the tumor core (active tumor along with its internal necrosis) from the surrounding vasogenic edema; while the second, applying "transfer learning," differentiates pure vasogenic edema from edema infiltrated by tumor cells. By combining these two pipelines and generating volumetric probability maps, it is possible to precisely delineate the proposed treatment area (tumor core plus infiltrative edema) and distinguish it from pure vasogenic edema. In addition to this information, the service provides an automated report containing spatial metrics for the resulting segmented areas, offering a robust tool for helping in the diagnosis and personalized treatment of this disease.

Keywords: glioblastoma, magnetic resonance imaging, microservices, deep learning, automated workflow.

Arquitectura de microservicios para el procesamiento automatizado y segmentación tridimensional avanzada del glioblastoma

Resumen. Este trabajo presenta el desarrollo de una plataforma escalable basada en microservicios en contenedores para el análisis automatizado de imágenes de resonancia magnética multiparamétrica de glioblastoma. El sistema integra una interfaz web que gestiona la ingesta dinámica de modalidades imagenológicas, realizando la transición del entorno clínico al computacional. El flujo de trabajo inicia con un preprocesamiento estandarizado de las imágenes, fundamentado en el protocolo de los desafíos BraTS. Este proceso incluye la conversión de formatos DICOM a NIfTI, la reorientación del sistema de coordenadas, y el registro rígido de la secuencia T1-Gd a un atlas anatómico de referencia bajo una resolución isotrópica de 1 mm^3. Posteriormente, las secuencias T1, T2 y T2-FLAIR son co-registradas a la T1-Gd y el tejido cerebral es extraído mediante máscaras generadas algorítmicamente. El núcleo analítico emplea redes neuronales profundas (transformers SwinUNETR seguidos de redes contrastivas) a través de dos pipelines de inferencia concurrentes. El primero discrimina el núcleo tumoral (tumor activo junto a su necrosis interna) del edema vasogénico circundante; mientras que el segundo, aplicando "transfer learning", diferencia el edema vasogénico puro del edema infiltrado por células tumorales. La conjunción de ambos flujos permite, a partir de la generación de mapas de probabilidad volumétricos, delimitar de forma precisa la zona de tratamiento propuesta (núcleo tumoral más edema infiltrado) y discriminarlo del edema vasogénico puro. Además de esta información, el servicio entrega un informe automatizado con métricas espaciales de las distintas zonas segmentadas, proporcionando una herramienta robusta de apoyo al diagnóstico y tratamiento personalizado de esta enfermedad.

Palabras clave: glioblastoma, resonancia magnética, microservicios, aprendizaje profundo, flujo automatizado.

1 Introducción

El glioblastoma (GBM) es el tumor cerebral primario más frecuente y agresivo en adultos. Una de sus características clínicas más desafiantes es su naturaleza altamente infiltrativa, donde las células tumorales migran varios centímetros desde la lesión primaria ocultándose dentro del edema peritumoral (Bobholz et al., 2024). Para definir adecuadamente la extensión de la resección quirúrgica o el área objetivo para la radioterapia, es imperativo distinguir el edema vasogénico puro del edema infiltrado. Sin embargo, esta infiltración microscópica es indetectable a simple vista tanto en secuencias estructurales convencionales (T1, T1 con contraste de gadolinio (T1Gd), T2 y T2-FLAIR) como en las funcionales de difusión o perfusión.

Recientemente, la integración de redes neuronales profundas con imágenes multiparamétricas ha demostrado una capacidad excepcional para capturar

características radiómicas sutiles asociadas a la infiltración. Arquitecturas avanzadas como los transformadores de visión SwinUNETR (Hatamizadeh et al., 2022) asociadas al aprendizaje contrastivo (Xu & Wong, 2025) y a una estrategia de aprendizaje por transferencia han logrado resultados prometedores en la segmentación tridimensional de estas subregiones complejas (Mulet de los Reyes et al., 2026; enviado para su publicación).

Sin embargo, trasladar estos complejos modelos matemáticos desde los entornos experimentales de laboratorio hacia la práctica clínica diaria representa un desafío tecnológico mayor (MLOps). Los modelos de inteligencia artificial (IA) médica son frágiles ante la variabilidad de los datos del mundo real: requieren imágenes preprocesadas con resoluciones isotrópicas estrictas, sin cráneo y perfectamente co-registradas (Bakas et al., 2022). Además, el procesamiento de volúmenes tridimensionales demanda un manejo eficiente de la memoria de las Unidades de Procesamiento Gráfico (GPU) y arquitecturas de software que eviten el bloqueo de los sistemas durante la inferencia.

El presente trabajo en desarrollo propone abordar esta brecha tecnológica mediante el diseño de un prototipo de servicio escalable basado en una arquitectura de microservicios. El objetivo es abstraer por completo la complejidad del preprocesamiento y la inferencia matemática, ofreciendo a la comunidad médica una herramienta automatizada y robusta que reciba imágenes crudas y entregue reportes clínicos procesables de ayuda en la planificación terapéutica del GBM.

2 Metodología

Para garantizar la escalabilidad, tolerancia a fallos e independencia de los entornos de ejecución, la plataforma fue diseñada utilizando un esquema de microservicios aislados y orquestados mediante contenedores (Docker). El sistema se divide en cuatro componentes lógicos principales conectados a través de volúmenes compartidos **(Figura 1).**

La interfaz de usuario o capa de presentación (Frontend) fue desarrollada para proporcionar una interacción intuitiva en entornos clínicos, permitiendo la carga dinámica de estudios en formato crudo (series DICOM) o estandarizado (NIfTI). Esta capa se comunica exclusivamente con una API REST (Backend) desarrollada en FastAPI (https://fastapi.tiangolo.com), la cual actúa como un puerto de enlace asíncrono. La API gestiona el almacenamiento inicial de los archivos, aísla las series imagenológicas y orquesta las peticiones del usuario sin bloquear el servidor web.

El análisis de imágenes médicas en 3D es computacionalmente intensivo. Para evitar tiempos de espera exhaustivos en el cliente (timeouts), se implementó un patrón de "cola de tareas" (task queue). El Backend delega las órdenes de ejecución a un gestor de mensajes (broker) basado en Redis (https://redis.io). Un microservicio independiente (worker), gestionado por Celery (https://docs.celeryq.dev/en/stable/) y con acceso directo al hardware acelerador (GPU), consume estas tareas en segundo plano e informa su progreso en tiempo real a la interfaz web mediante sondeo (polling).

Para compatibilizar los estudios clínicos del mundo real con el espacio vectorial esperado por los modelos entrenados bajo protocolos estrictos, el worker

ejecuta un pipeline de estandarización completamente automatizado utilizando la librería matemática ANTsPy (https://github.com/ANTsX/ANTsPy).

El flujo de pre-procesamiento automatizado consiste en: 1) Detección y conversión de series DICOM a volúmenes NIfTI utilizando dicom2nifti. 2) Corrección de inhomogeneidades del campo magnético. 3) Reorientación al sistema de coordenadas LPS (izquierda, posterior, superior). 4) Registro rígido de la imagen T1-Gd al atlas anatómico de referencia del Stanford Research Institute (SRI24). 5) Co-registro del resto de las modalidades (T1, T2, T2-FLAIR, modalidades de difusión y perfusión) hacia la T1-Gd ya registrada. 6) Aplicación de la transformación acumulada para llevar todas las secuencias a una resolución isotrópica de 1 mm^3. 7) Extracción del cráneo mediante la multiplicación algorítmica de los volúmenes por la máscara de tejidos del atlas utilizado (https://www.nitrc.org/projects/sri24).

El motor de inferencia o núcleo analítico, desarrollado en PyTorch (*https://pytorch.org)* y MONAI (https://project-monai.github.io/), reside en la memoria de video del worker. El mismo consta de dos pipelines (arquitecturas SwinUNETR independientes acopladas a cabeceras de proyección contrastiva). Para procesar el masivo volumen resultante de ensamblar las 11 secuencias imagenológicas sin exceder los límites físicos de la VRAM, se implementó una estrategia de inferencia por ventana deslizante (sliding window inference), procesando parches de 128x128x128 vóxeles con un solapamiento del 25%.

Una vez generados los mapas de probabilidad derivados de cada pipeline, el sistema los combina lógicamente y calcula los volúmenes absolutos (V) extrayendo las dimensiones físicas espaciales de la cabecera NIfTI (Sx, Sy, Sz):

$$V = \frac{Nvoxels \times (Sx, Sy, Sz)}{1000}$$

Finalmente, un módulo generador de reportes renderiza un archivo PDF que incluye las métricas clave de rendimiento volumétrico para cada segmentación ejecutada junto a una composición visual multipanel de los mapas de calor (Jet colormap) centrada automáticamente en el corte axial de mayor área tumoral.

3 Resultados preliminares

El prototipo ha sido desplegado exitosamente en un entorno de desarrollo equipado con aceleración por hardware convencional (NVIDIA RTX 3080 Ti). Las pruebas de estrés end-to-end demostraron que la arquitectura asíncrona es capaz de recibir series DICOM crudas y gestionar el pipeline completo sin interrupciones. El microservicio de preprocesamiento ANTsPy logró estandarizar, co-registrar y extraer el cráneo de 11 modalidades distintas en un tiempo promedio de 3.5 minutos. Por su parte, la tarea de inferencia volumétrica por ventana deslizante cruzando los dos modelos SwinUNETR completó la segmentación total del cerebro en aproximadamente 22 segundos. La **tabla 1** muestra las principales métricas de exactitud diagnóstica obtenidas al testear los dos modelos SwinUNETR completos.

Desde el punto de vista clínico, el sistema logra abstraer la salida del modelo de IA profundo, entregando al especialista un reporte PDF (**Figura 2**) que analiza la

región de interés discriminando zonas de alto valor para una eventual cirugía y/o radioterapia (zona de tratamiento propuesta, que incluye el núcleo tumoral y el edema infiltrado periférico) respecto del edema puramente vasogénico no infiltrado, garantizando un flujo de trabajo utilizable y de alto rendimiento. Los mapas de probabilidades y las segmentaciones generadas son exportadas en volúmenes NIfTI, para su análisis y renderizado en cualquier visualizador de imágenes médicas como el ITK-SNAP (https://www.itksnap.org/pmwiki/pmwiki.php) (**Figura 3**).

Métrica	Zona de tratamiento	Edema vasogénico	Cerebro normal	Global
Dice	0.697 ± 0.09	0.650 ± 0.13	0.996 ± 0.001	0,781 ± 0,07
Sensibilidad	0.732 ± 0.17	0.864 ± 0.06	0.993 ± 0.001	0.863 ± 0.08
Precisión	0.692 ± 0.09	0.553 ± 0.18	0.999 ±0.0003	0.748 ± 0.09
AUC-ROC	0.871 ± 0.08	0.869 ± 0.09		0.870 ± 0.08
Exactitud				0.794 ± 0.06

Table 1. Métricas correspondientes al modelo completo.

4 Conclusiones y Trabajo futuro

La transición de modelos experimentales de inteligencia artificial hacia herramientas de uso clínico requiere de una ingeniería de software robusta. La arquitectura de microservicios presentada demuestra ser una solución viable, escalable y eficiente para orquestar pipelines de neuroimágenes complejos. Al automatizar el riguroso preprocesamiento espacial y utilizar colas de tareas asíncronas para la inferencia de modelos pesados (SwinUNETR), el sistema elimina barreras técnicas significativas para el personal médico.

Las acciones futuras sobre esta línea de desarrollo se centran en la preparación del sistema para entornos de producción hospitalarios. Esto incluye la orquestación del despliegue en la nube mediante Kubernetes (https://kubernetes.io/), la integración de un nodo receptor DICOM nativo (compatibilidad directa con sistemas PACS, https://visualmedica.com/que-es-pacs-y-como-funciona/), y la ejecución de ensayos de validación clínica multicéntrica que permitan certificar la utilidad del reporte automatizado en la planificación terapéutica real de pacientes con glioblastoma.

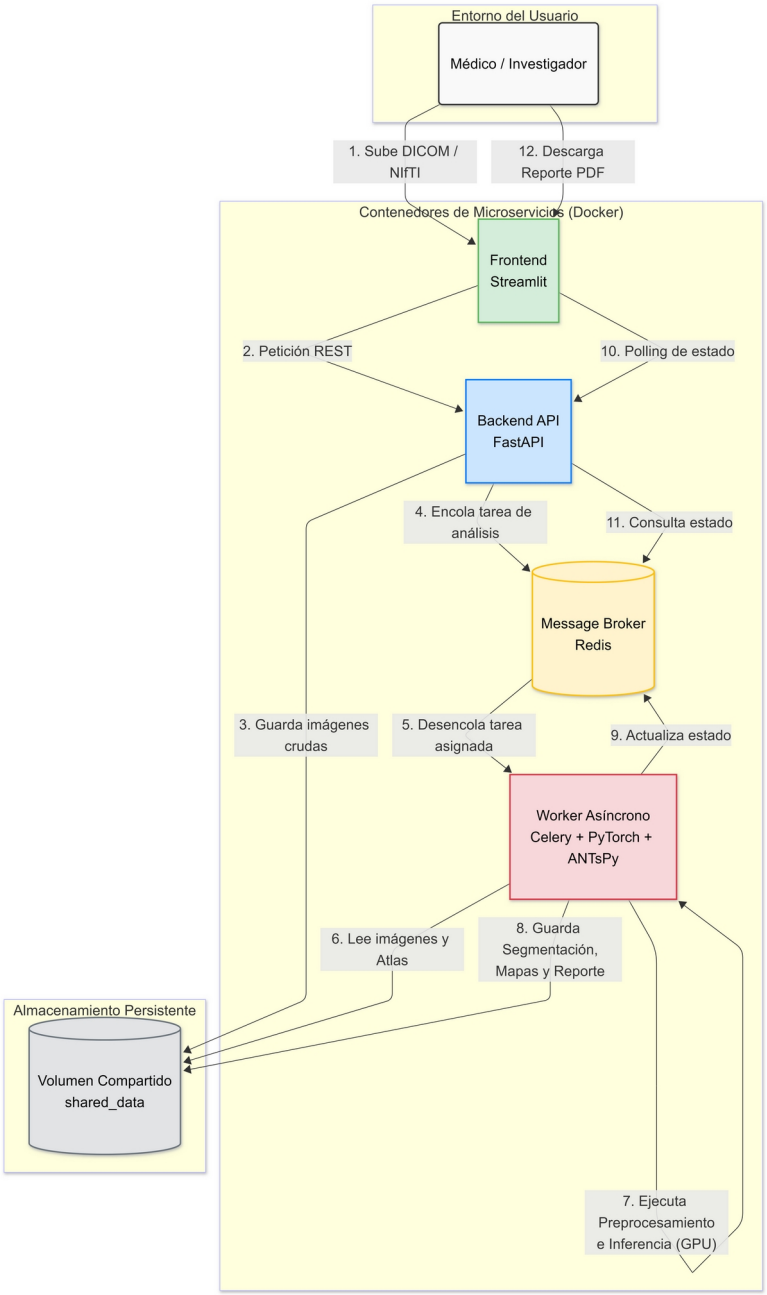

Figura 1. Diagrama de la arquitectura de microservicios del sistema. Se observan los contenedores independientes (Frontend, Backend, Broker y Worker) y el flujo asíncrono de los datos DICOM/NIfTI a través de los volúmenes compartidos.

Reporte de Análisis Oncológico Multi-Pipeline

ID del Caso Clínico: caso0086

Corte Analizado (Plano Axial): Z = 90

Volumetría Detallada

Región / Componente	Volumen (cm3)
Tumor Core (Necrosis + Activo)	58.69 cm3
Infiltración Tumoral	15.45 cm3
Zona de Tratamiento Total (Core + Infiltración)	**75.23 cm3**

Visualización Multipanel (Escala de Calor Jet)

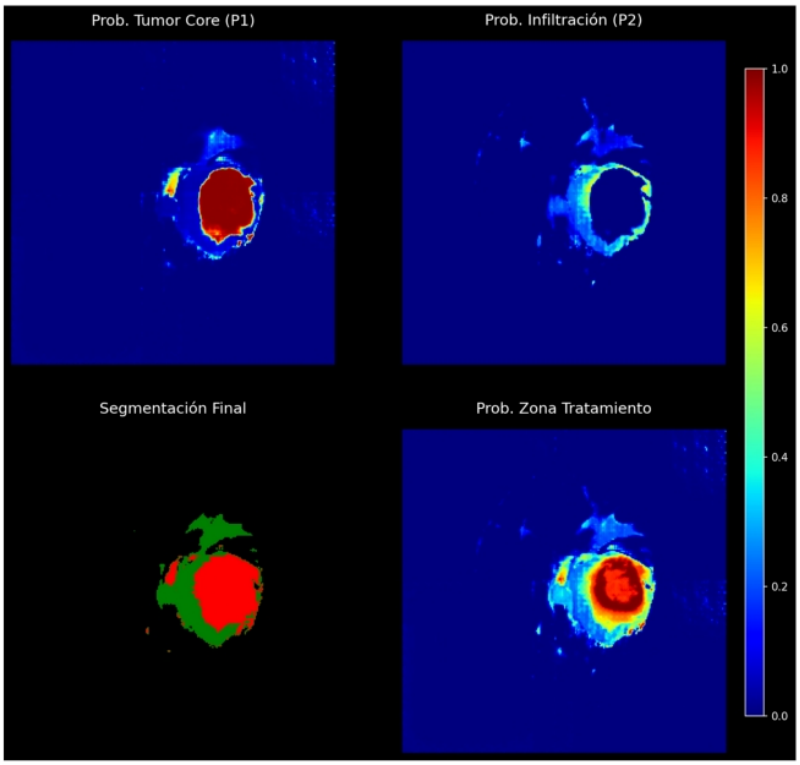

Fig. 2. Ejemplo del reporte clínico automatizado generado por el sistema. Se observan los volúmenes de las principales regiones junto a los mapas de probabilidad renderizados.

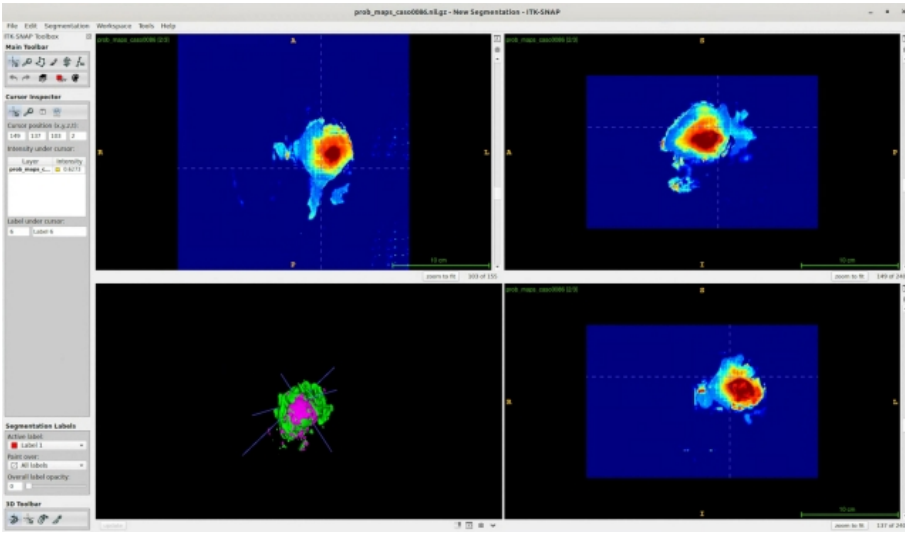

Fig. 3. Ejemplo de carga en el software ITK-SNAP del mapa de probabilidad tumoral generado por el sistema para la zona de tratamiento.

Referencias

Bakas, S., Sako, C., Akbari, H., Bilello, M., Sotiras, A., Shukla, G., et al. (2022). The University of Pennsylvania glioblastoma (UPenn-GBM) cohort: Advanced MRI, clinical, genomics & radiomics. Scientific Data, 9, 453

Bobholz S, Lowman A, Connelly J, Duenweg S, Winiarz A, Nath B, et al. (2024). Noninvasive autopsy-validated tumor probability maps identify glioma invasion beyond contrast enhancement. Neurosurgery, 95 (3): 537-547

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., & Xu, D. (2022). Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. ArXiv, 2201.01266.

Taha, A. & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: Analysis, selection and tool. BMC Medical Imaging, 15, 29

Xu X & Wong S. (2025). Contrastive learning in brain imaging. Computerized Medical Imaging and Graphics (2025), 121, 102500