UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingenieros Informáticos

Máster Universitario en Inteligencia Artificial (MUIA)

Optimización Radiómica y Clasificación SLMVP-SVM Explicable para el Diagnóstico Diferencial de Abscesos y Glioblastomas en Tomografía Computarizada

T R A B A J O D E F I N D E M Á S T E R

Presentado para la obtención del título de Máster por:

Diego Gálvez López

Bajo la supervisión de:
Dr. Esteban García Cuesta

Madrid, 2026

Agradecimientos

Este trabajo no hubiese sido posible sin la ayuda de dos profesores:

En primer lugar Esteban García Cuesta, que me ha ofrecido su disponibilidad y tiempo en el desarrollo del mismo a la par que me ha guiado en el enfoque y los errores que han ido surgiendo a lo largo del trabajo. Haber finalizado este máster con él como profesor y luego como tutor de mi TFM ha sido especialmente valioso.

En segundo lugar a Carmen Sánchez Ávila, sin quien probablemente no hubiese encontrado la motivación para hacer este máster y cuyo acompañamiento ha sido determinante en mi desarrollo académico y profesional. Quiero expresar mi más profundo agradecimiento por todo lo que ha hecho por mí estos últimos años.

Resumen

Los abscesos cerebrales (AC) y los glioblastomas (GBM) presentan hallazgos radiológicos similares, lo que dificulta su diagnóstico diferencial en la práctica clínica. En este Trabajo Fin de Máster se desarrolla un enfoque radiómico aplicado a imágenes de tomografía computarizada (TC), con el objetivo de proporcionar una muestra de estudio representativa y optimizar los algoritmos de clasificación empleados en la bibliografía para la discriminación automática entre ambas entidades.

Se recopilaron un total de 62 casos del Hospital Universitario Ramón y Cajal, donde las lesiones fueron segmentadas manualmente e interpoladas a 1 mm para extraer 107 características radiómicas (forma, primer orden, GLCM, GLDM, GLRLM, GLSZM, NGTDM) utilizando Synapse3D y 3DSlicer.

Se evaluaron cinco clasificadores: tres de ellos basados en árboles de decisión (Random Forest, XGBoost y Explainable Boosting Machine (EBM) así como una SVM) y un nuevo modelo que aplica SVM junto con una proyección previa de características (SLMVP-SVM), optimizado mediante búsqueda en malla y validación cruzada estratificada. Además, se integró un análisis explicativo personalizado con SHAP para el modelo SVM, lo que permitió identificar las características radiómicas más relevantes en la diferenciación de AC y GBM.

Los resultados muestran que el modelo SLMVP–SVM reentrenado con 8 características (rank = 8) alcanzó el mejor rendimiento medio en test, con *Accuracy*=0.7719 ± 0.0736, *F1*=0.7396 ± 0.0941, *Recall*=0.6741 ± 0.1386, *Precision*=0.8568 ± 0.1182 y AUC=0.8007 ± 0.0714, mejorando la exactitud frente a los modelos estándar (accuracy $\approx 0.72-0.73$) y manteniendo un compromiso favorable entre rendimiento y parsimonia.

Este trabajo refuerza la viabilidad de la TC como herramienta rápida y accesible para el diagnóstico diferencial inicial, y sienta las bases para integrar modelos radiómicos explicables en la rutina clínica, especialmente cuando se combinen con segmentaciones automáticas y datos clínicos en futuros estudios.

Abstract

Brain abscesses (BA) and glioblastomas (GBM) present similar radiological findings, which makes their differential diagnosis challenging in clinical practice. This Master's Thesis develops a radiomic approach applied to computed tomography (CT) images, with the aim of providing a representative study sample and optimizing the classification algorithms reported in the literature for the automatic discrimination between both entities.

A total of 62 cases were collected from the Ramón y Cajal University Hospital, where the lesions were manually segmented and interpolated to 1 mm in order to extract 107 radiomic features (shape, first order, GLCM, GLDM, GLRLM, GLSZM, NGTDM) using Synapse3D and 3DSlicer.

Five classifiers were evaluated: three based on decision trees (Random Forest, XGBoost, and Explainable Boosting Machine (EBM)), a Support Vector Machine (SVM), and a new model that combines SVM with a prior feature projection (SLMVP-SVM), optimized through grid search and stratified cross-validation. In addition, a customized SHAP-based explanatory analysis was integrated for the SVM model, enabling the identification of the most relevant radiomic features for differentiating BA and GBM.

Results show that the retrained SLMVP–SVM model using 8 features (rank = 8) achieved the best mean test performance, with *Accuracy*=0.7719 ± 0.0736, *F1*=0.7396 ± 0.0941, *Recall*=0.6741 ± 0.1386, *Precision*=0.8568 ± 0.1182, and AUC=0.8007 ± 0.0714, improving overall accuracy compared with the standard models (accuracy $\approx 0.72-0.73$) while preserving a compact and more interpretable feature set.

This work reinforces the feasibility of CT as a rapid and accessible tool for initial differential diagnosis and lays the groundwork for integrating explainable radiomic models into clinical routine, particularly when combined with automatic segmentations and clinical data in future studies.

Palabras Clave

Absceso cerebral, Algoritmo, Diagnóstico, Detección automática, EBM, Glioblastoma, Machine learning, Random Forest, Radiómica, Resonancia magnética, SVM, Segmentación, Selección de características, SHAP, Tomografía computarizada, Neurología, Neuroimagen, XGBoost.

Keywords

Brain abscess, Algorithm, Diagnosis, Automatic detection, EBM, Glioblastoma, Machine learning, Random Forest, Radiomics, Magnetic resonance imaging, SVM, Segmentation, Feature selection, SHAP, Computed tomography, Neurology, Neuroimaging, XGBoost.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción y objetivos

1.0.1 Introducción

El **glioblastoma (GBM)** [28] es un tipo de tumor del sistema nervioso central (SNC) que se forma a partir del tejido glial (de sostén) del encéfalo y la médula espinal; tiene células cuyo aspecto es muy diferente al de las células normales (Figura 1.1). Por lo general, el glioblastoma se presenta en adultos y afecta más al encéfalo (cerebro) que a la médula espinal. También se llama astrocitoma de grado IV, GBM, glioblastoma multiforme y glioma maligno.

Es el tumor más frecuente entre tumores primarios malignos del sistema nervioso central y se caracteriza por ser muy complicado de tratar debido a su crecimiento acelerado. Puede presentarse en cualquier etapa de la vida, aunque es más habitual en varones y en personas de edad avanzada. Entre sus manifestaciones clínicas se encuentran la cefalea progresiva, las náuseas y vómitos, los trastornos visuales (como visión borrosa o doble) y la aparición de crisis epilépticas.

Actualmente, la resección quirúrgica es el tratamiento estándar para esta enfermedad, seguido de un tratamiento de radioterapia y quimioterapia postoperatorias.

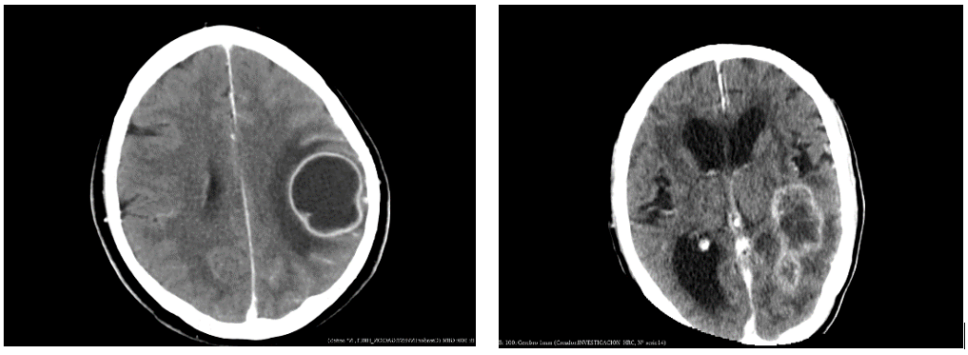


Figura 1.1: Imágenes de TC de dos glioblastomas. El de la izquierda, más uniforme y con los bordes definidos, el de la derecha presenta márgenes gruesos irregulares

Por su parte, el **absceso cerebral (AC)** [25] se define como una acumulación de pus en el parénquima cerebral, originada a partir de distintos focos infecciosos. Puede desarrollarse secundariamente a procesos locales como osteomielitis de los huesos craneales, sinusitis o mastoiditis, así como tras traumatismos penetrantes del cráneo (Figura 1.2). Asimismo,

puede aparecer por diseminación hematógena, cuando bacterias u otros patógenos alcanzan el sistema nervioso central a través del torrente sanguíneo desde infecciones situadas en otros órganos.

La sintomatología deriva principalmente del aumento de la presión intracraneana asociado al efecto de masa (compresión y desplazamiento de las estructuras cerebrales vecinas por el incremento de volumen de la lesión [36]) y, en ocasiones, de la propia lesión cerebral focal. Entre los síntomas más frecuentes se incluyen cefalea, náuseas, vómitos, somnolencia, cambios de personalidad y déficits neurológicos focales, que suelen instaurarse de forma progresiva a lo largo de varios días o semanas; no obstante, en algunos pacientes estas manifestaciones pueden ser discretas o incluso ausentes hasta fases avanzadas de la enfermedad.

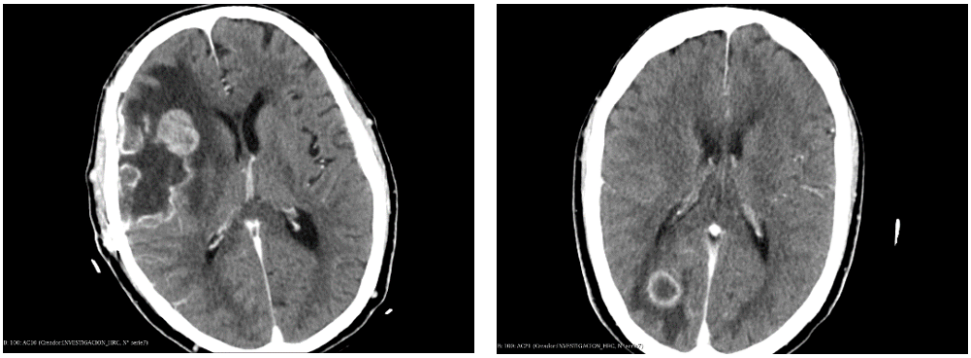


Figura 1.2: Imágenes de TC de dos abscesos cerebrales. El de la izquierda, más heterogéneo e irregular, con un anillo de espesor no uniforme, el de la derecha, de menor tamaño y más regular.

Su tratamiento consiste en una administración de antibióticos durante un mínimo de 4 a 8 semanas de ceftriaxona o cefotaxima. Además, en la mayoría de los abscesos es necesario el drenaje estereotáctico guiado por TC o a cielo abierto, sobre todo en aquellos de más de 2cm de diámetro.

Diferencias entre ambas lesiones

Hay diferencias fundamentales entre las características biológicas, las opciones terapéuticas y el pronóstico del absceso cerebral y del glioblastoma necrótico. En el caso del GBM, el pronóstico suele ser desfavorable, con una mediana de supervivencia de aproximadamente 15 meses tras el diagnóstico y una supervivencia global (SG) en torno a 22 meses [39]. En contraste, la mortalidad asociada al absceso cerebral ha disminuido de forma considerable en las últimas décadas. Cuando el tratamiento estándar se instaura de manera precoz, muchos pacientes pueden evitar intervenciones quirúrgicas de alto riesgo y conseguir una recuperación postoperatoria más favorable.

La tomografía computarizada y la resonancia magnética (MRI) constituyen herramientas clave en el diagnóstico diferencial entre abscesos cerebrales y glioblastomas. Ambas técnicas permiten identificar rasgos de imagen característicos que facilitan la distinción entre estas entidades. La TC se emplea con frecuencia como estudio inicial debido a su rapidez, amplia disponibilidad y capacidad para evidenciar lesiones con efecto de masa, calcificaciones o hemorragias.

Se ha comprobado que el uso combinado de TC y RM mejora la precisión diagnóstica,

permitiendo diferenciar entre lesiones infecciosas y neoplásicas, lo cual es crucial para establecer un tratamiento adecuado [4].

Concretamente, la tomografía computarizada (CT) es una modalidad de imagen médica que usa equipo especial de rayos X para crear imágenes detalladas, o exploraciones, de regiones internas del cuerpo. A veces, también se llama tomografía axial computarizada (TAC).

Los equipos actuales de tomografía computarizada (CT) [27] adquieren los datos de forma continua siguiendo una trayectoria helicoidal (o espiral), en lugar de registrar secuencialmente cortes aislados del cuerpo como hacían los primeros dispositivos. Este modo de adquisición aporta varias ventajas frente a las técnicas convencionales: reduce el tiempo de exploración, permite reconstruir volúmenes tridimensionales de mayor calidad de las regiones internas y mejora la detección de lesiones de pequeño tamaño.

En una TC helicoidal, el paciente permanece tumbado e inmóvil sobre una camilla que se desplaza lentamente a través del gantry o "rosquilla" del escáner, mientras el tubo de rayos X gira emitiendo un haz en forma de abanico. La radiación que atraviesa el cuerpo y no es absorbida es registrada por una matriz de detectores, y a partir de estas señales se reconstruyen imágenes bidimensionales del órgano de interés (en este caso, el cerebro) en múltiples cortes, cuyo grosor puede variar desde fracciones de milímetro hasta aproximadamente 4–5 mm, según las prestaciones del equipo.

Con frecuencia, para mejorar la delimitación de estructuras y lesiones, se recurre a la administración de un *medio de contraste*. Este contraste, que suele administrarse por vía intravenosa (o, en otras exploraciones, por vía oral o rectal), resalta determinadas regiones internas y genera imágenes más nítidas. Los compuestos yodados y las suspensiones de bario son los agentes de contraste utilizados con mayor frecuencia en las exploraciones de tomografía computarizada.

La duración de una tomografía computarizada helicoidal multi slice cerebral de última generación es capaz de tomar 128 cortes en aproximadamente 45 segundos [17] con un alcance de exploración que puede llegar a los 0,5 mm.

La principal ventaja de la tomografía computarizada frente al uso de resonancia magnética (RM) para la detección de estas patologías cerebrales radica principalmente en un criterio de disponibilidad. La duración de un estudio de resonancia magnética cerebral suele ser de aproximadamente 30 minutos, unas 30 veces más que un estudio completo de TC de la misma región. Además, los equipos que se utilizan en las pruebas de TC suelen ser más económicos que los utilizados en la resonancia magnética por lo que suele haber más disponibilidad y número de ellos en sistemas hospitalarios.

Estas características, además, hacen que TC sea la primera prueba a realizar en el servicio de urgencias para el diagnóstico.

Generalmente, en tomografía computarizada con contraste la diferencia entre ambos tipos de lesiones suele ser relativamente clara.

Los glioblastomas suelen ser tumores grandes en el momento del diagnóstico [32]. A menudo tienen márgenes gruesos que se realzan irregularmente y un núcleo necrótico central, que también puede tener un componente hemorrágico. Están rodeados de edema de tipo vasogénico, que de hecho suele contener infiltración de células neoplásicas. En TC sus principales características son (Figura 1.3):

- Efecto de masa marcado
- Márgenes gruesos irregulares hiperatenuantes (alta celularidad)
- Centro hipodenso irregular que representa la necrosis
- Edema vasogénico circundante
- Casi siempre está presente un intenso realce irregular y heterogéneo de los márgenes

Por otra parte, los abscesos cerebrales en TC suelen presentar las siguientes características (Figura 3) [31]:

- Borde hipodenso exterior e hiperdenso interior (signo de doble llanta) en la mayoría de los casos.
- Anillo de tejido isodenso o hiperdenso, típicamente de espesor uniforme
- Baja atenuación central (líquido/pus)
- Baja densidad circundante (edema vasogénico)
- Puede haber ventriculitis, que se observa como un realce del epéndimo

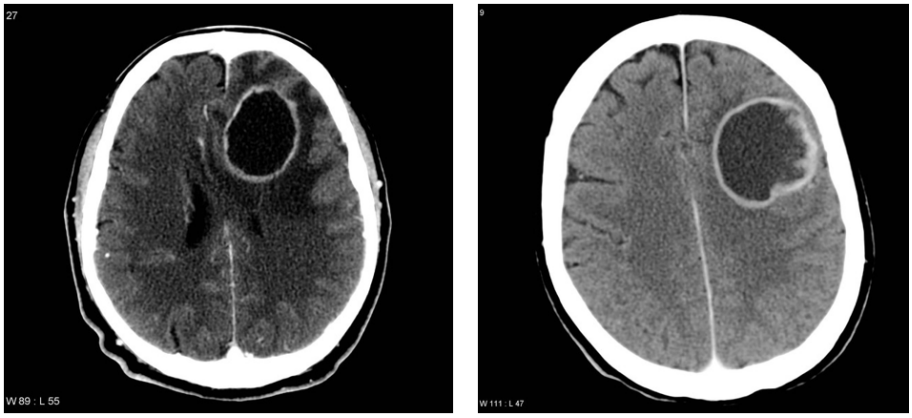


Figura 1.3: Imágenes de TC de un absceso con un anillo hiperdenso y signo de doble llanta (izquierda) y un glioblastoma con necrosis y realce irregular (derecha).

No obstante, existen rasgos compartidos por ambos tipos de lesiones que en algunas situaciones pueden hacer más compleja la valoración del radiólogo:

- **Realce anular tras la administración de contraste**: Tanto los abscesos como los tumores muestran un patrón de realce en anillo secundario a la ruptura de la barrera hematoencefálica, lo que puede dificultar una discriminación inicial en estudios contrastados de TC o RM [35].
- **Edema perilesional**: Ambos procesos originan un edema vasogénico importante, que se manifiesta como áreas hipodensas en TC e hiperintensas en T2/FLAIR. Este edema contribuye al efecto de masa y se asocia a síntomas neurológicos como déficits

focales y crisis epilépticas [35].

- **Restricción de la difusión en DWI**: Los abscesos suelen presentar una restricción intensa de la difusión por la elevada viscosidad del pus; sin embargo, un patrón similar puede observarse en glioblastomas necróticos, especialmente en zonas de necrosis central compacta [41].

- **Valores bajos de ADC**: Los abscesos cerebrales suelen mostrar un ADC reducido debido a la elevada viscosidad del contenido purulento, que restringe el movimiento de las moléculas de agua. No obstante, [10] describen que los glioblastomas necróticos también pueden presentar áreas con ADC bajo en regiones de necrosis densa.

Aunque estas características pueden ser reconocidas visualmente por el especialista, las imágenes de TC contienen una gran cantidad de información cuantitativa que no resulta directamente accesible al ojo humano sin un procesamiento adicional. El análisis radiómico permite explotar esa información latente y extraer descriptores más sutiles de la lesión, que aportan valor añadido para su correcta diferenciación.

La **radiómica** [34] es una ciencia que de manera no invasiva estudia características de las imágenes médicas imperceptibles al ojo humano mediante la aplicación de algoritmos automatizados, con el objetivo de asociarlas a estados fisiológicos concretos.

Estas características se dividen en las siguientes categorías: características de forma, de primer orden, matriz de co-ocurrencia de niveles de gris (GLCM), matriz de dependencia de niveles de gris (GLDM), matriz de longitud de carrera de niveles de gris (GLRLM), matriz de tamaño de zona de niveles de gris (GLSZM) y matriz de diferencia de tono gris vecino (NGTDM).

Esta ciencia ómica se presenta como una herramienta de apoyo en la investigación y en la práctica clínica, ofreciendo múltiples aplicaciones en áreas como la oncología, las enfermedades reumatológicas o las neurodegenerativas.

1.0.2 Objetivos

El objetivo de este TFM es doble: Probar cuatro clasificadores clásicos basados en métodos de machine learning sobre el conjunto de datos (EBM, Random Forest, XGBoost y SVM) y comparar sus métricas así como sus explicaciones y en segundo lugar incluir un nuevo clasificador con proyección previa de características, SLMVP-SVM junto con una implementación propia de SHAP que permita aproximar la importancia de cada una de las características en la toma de decisión. Para lograr esto se realizarán los siguientes pasos:

1. **Obtención del dataset**

   Obtener una base de datos balanceada de muestras abscesos y glioblastomas previamente seleccionados por el radiólogo junto con sus características radiómicas, tras un proceso de segmentado de las lesiones y extracción.

2. **Implementación y evaluación de los clasificadores.**

   Implementar cuatro clasificadores de *machine learning*: Explainable Boosting Machine (EBM), Random Forest (RF), XGBoost (XGB) y Support Vector Machine (SVM)

sobre el *dataset*, utilizando un enfoque *Leave-One-Out* en el conjunto de *train* a lo largo de varias iteraciones, con el objetivo de obtener métricas de evaluación fiables y comparables entre sí.

Analizar la explicabilidad ofrecida por *SHAP* en cada modelo, identificando qué características contribuyen más a la detección de la lesión y contrastando dichos resultados con el conocimiento clínico existente sobre las mismas.

3. **Aplicación de SLMVP-SVM sobre el dataset**

   Incorporar un flujo de preprocesamiento basado en la proyección supervisada Supervised Local Maximum Variance Preserving (SLMVP) para reducir la dimensionalidad del dataset y optimizar el espacio de características antes del entrenamiento con SVM. Evaluar el impacto de esta proyección en las métricas de clasificación.

   Comparar estos resultados con los obtenidos con los modelos clásicos y con SVM sin la proyección previa.

4. **Implementación de SHAP en el nuevo modelo**

   Integrar SHAP en SVM–SLMVP para estimar contribuciones por característica, generando rankings globales y explicaciones locales, y validando la coherencia clínica de las variables más influyentes con los clasificadores previamente descritos.

5. **Reentrenamiento del mejor modelo**

   Utilizar el análisis de explicabilidad de los distintos modelos para obtener las características más relevantes y reentrenar el mejor modelo usando sólo esas características, con el fin de hacer un modelo altamente interpretable en la práctica clínica.

6. **Evaluación final**

   Comparar las distintas métricas de evaluación del nuevo modelo con los anteriores, así como su explicabilidad y potencialidad de aplicación.

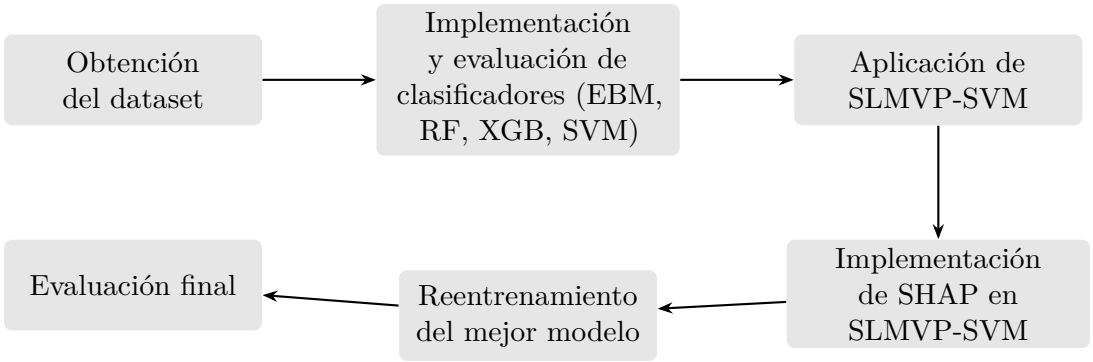


Figura 1.4: Secuencia de pasos del proyecto

Capítulo 2

Estado del Arte

El diagnóstico diferencial entre absceso cerebral (AC) y glioblastoma (GBM) ha sido un reto clínico en los últimos años, dado que, como hemos mencionado, ambas lesiones presentan características morfológicas similares, como el realce anular, el edema vasogénico tanto en resonancia magnética como en tomografía computarizada. Sin embargo, las diferencias en su estructura microscópica y composición tisular han motivado el desarrollo de técnicas avanzadas de imagen y de análisis cuantitativo para su discriminación.

2.0.1 Evolución histórica de la investigación

Los primeros estudios orientados al diagnóstico diferencial entre abscesos cerebrales (AC) y glioblastomas (GBM), desarrollados entre finales de los años noventa y la primera década de los 2000, se centraron en la caracterización física y bioquímica de la señal mediante técnicas de imagen funcional como la resonancia magnética ponderada por difusión (DWI), la espectroscopia por resonancia magnética de protón (^{1}H-MRS) y la imagen ponderada por susceptibilidad magnética (SWI). En el ámbito de la difusión, B.Desprechins y K.Stadnik [11] demostraron que los abscesos presentan una marcada restricción de difusión, con valores de coeficiente de difusión aparente (ADC) notablemente bajos ($\approx 0{,}21$–$0{,}34 \times 10^{-3}\,\mathrm{mm^2/s}$), frente a los tumores necróticos, donde el ADC es elevado ($\approx 2{,}2 \times 10^{-3}\,\mathrm{mm^2/s}$).

Por su parte, Kadota et al. [20] evidenciaron que la espectroscopia ^{1}H-MRS permite una discriminación metabólica precisa entre ambas lesiones gracias al metabolismo bacteriano en los abscesos.

Toh et al. [40] aplicaron la técnica SWI para analizar la morfología del borde de la lesión, identificando el denominado *dual rim sign* (un anillo doble con borde externo hipointenso e interno hiperintenso) como marcador altamente específico del absceso (precisión = 90.6 %). Este signo, junto con la presencia de bordes completos y lisos, permite diferenciar abscesos de GBM en RM incluso en casos donde la difusión no resulta concluyente.

A partir de 2020, el desarrollo de la radiómica y el aprendizaje automático introdujo una nueva etapa basada en la cuantificación masiva de características (forma, textura, intensidad, heterogeneidad) y su clasificación mediante diversos modelos como Random Forest, SVM, XGBoost o LightGBM. Los trabajos más recientes han alcanzado métricas de AUC en torno a 0.88–0.91, centrados casi exclusivamente en RM (T1CE, T2, FLAIR, DWI). Sin embargo, la aplicación de esta metodología a TC sigue siendo anecdótica, pese

a su mayor disponibilidad y rapidez diagnóstica en contexto de urgencias.

En este contexto, el presente TFM constituye una extensión natural de la línea de investigación actual, proponiendo por primera vez un pipeline radiómico explicable sobre CT con un número amplio de casos y la incorporación de un modelo propio optimizado, acompañado de su correspondiente análisis de explicabilidad.

2.0.2 Síntesis de estudios previos

El **Cuadro 2.1** presenta los trabajos más relevantes que abordan la diferenciación entre abscesos cerebrales y tumores (principalmente glioblastomas) así como la caracterización individualizada de varios tipos de tumor, agrupados según la técnica empleada y su desempeño cuantitativo. Cabe destacar que la mayoría de referencias encontradas no responden a la diferenciación directa entre los dos tipos de lesiones a excepción de [44], [2] y [38]. Además, el uso de tomografía computarizada para el estudio de este tipo de lesiones no es común, dado que la resonancia magnética (MRI) es preferible para la evaluación de tejidos blandos y la caracterización microestructural de las lesiones, pese a su mayor tiempo de adquisición de imagen, por ello la mayoría de los trabajos aquí expuestos están realizados sobre esta modalidad de imagen.

Referencia	Año	Modalidad / Secuencias	Población	Modelo / Métricas	Hallazgos principales
Diferenciación entre GBM y AC					
Bo et al., *Frontiers in Medicine*	2021	MRI (T1WI, T2WI)	188 (102 AC, 86 CG)	SVM-RFE, AUC=0.85	Fusión HCR+DTL (T2WI-DLR); distingue AC vs GC con 85 % AUC y alta reproducibilidad (ICC>0.95).
Xiao et al., *J. Integr. Neurosci.*	2021	MRI (CE-T1WI, T2-FLAIR)	118 (86 GBM, 32 AC)	LR, RF; AUC=0.89–0.99	ROI total + relación edema/tumor optimizan diagnóstico; modelo combinado AUC ≈ 0.97 (GBM: heterogéneo, AC: VR bajo).
Solar et al., *Sci. Rep.*	2023	MRI (DWI / ADC maps)	40 (26 GBM, 14 AC)	kNN, SVM; Acc=0.85–0.90	Análisis lineal ADC (LOI) sin ROI fija; gradientes edema–anillo distinguen GBM vs AC (90 %).
Diferenciación entre GBM y metástasis cerebral (BM)					
Priya et al., *Sci. Rep.*	2021	MRI (T1W, T2W, T1-CE, ADC, FLAIR; 1.5 T Siemens)	120 intracraneales (60 GBM, 60 metástasis)	LASSO, ENet, AdaBoost; AUC=0.95, Acc=0.89	Radiomics multiparamétrico distingue GBM vs. metástasis con alta precisión; la esfericidad y edema fueron las variables más discriminantes.
Liu et al., *Front. Neurosci.*	2022	MRI (T2WI, CE-T1WI)	935 (train+val ext.)	TPOT AutoML; AUC=0.99 (int.), 0.87 (ext.)	Diferenciación GBM/BM; T2 >CE; rendimiento superior a radiólogos (AUC 0.51–0.63).
Pomohaci et al., *Diagnostics*	2025	MRI (T2W-FLAIR, CE-T1W)	78 (39 GBM, 39 BM)	CNN 3D (U-Net)	Segmentación automática de volúmenes; GBM con mayor necrosis y BM con más edema; resultados equivalentes al método manual.
Xia et al., *Acad. Radiol.*	2025	MRI (CE-T1WI)	434 (226 GB, 208 BM)	LGBM, SHAP; AUC=0.92	Features multiescala (LoG, wavelet) captan heterogeneidad GB vs homogeneidad BM; interpretabilidad clínica vía SHAP.
Caracterización individual de lesiones					
Meaney et al., *J. Theor. Biol.*	2023	MRI (T1WI, T1-GAD, T2WI, T2-FLAIR, DWI, ADC)	5 (GBM grado IV) + 2500 tumores sintéticos	PINN basado en modelo PI; Acc=0.90	La red estima difusividad (D) y proliferación (r) a partir de DWI/ADC en dos tiempos; error <5 % en 89.7 % de casos; predice progresión y mapas de celularidad.
Revisiones y síntesis radiómicas					
Yi et al., *Front. Oncol.*	2021	MRI/PET multimodal	≈ 40 estudios	CNN, RF, LASSO, Cox, AUC=0.78–0.99	Alto potencial diagnóstico y pronóstico. Limitaciones: reproducibilidad y estandarización.
Cerrone et al., *Cancers*	2022	MRI (CE-T1, T2, FLAIR, ADC)	42 estudios (2015–2022)	SVM, LR, RF, Radiomics logra AUC=0.7–0.99	Alto potencial diagnóstico pero pobre validación externa.

Tabla 2.1: Principales estudios sobre caracterización y diferenciación entre abscesos cerebrales y tumores (GBM y otros), agrupados por técnica de imagen y enfoque metodológico.

Observamos que la investigación en este ámbito ha evolucionado desde enfoques cualitativos

y basados en la experiencia radiológica hacia metodologías cuantitativas y automatizadas apoyadas en la radiómica y el aprendizaje automático. A pesar de los notables avances en MRI, la mayoría de los trabajos presentan limitaciones comunes:

1. **Muestras reducidas y escasa validación externa**: Observamos que la media de muestras por estudio es de 149 ± 214, pero si nos enfocamos solo en aquellos que comparan directamente entre abscesos y glioblastomas esta media es mucho más pequeña, 72 ± 78, ambas con una gran desviación indicando la variabilidad entre el número de pacientes utilizados en cada estudio. Además en estos casos no se realizó validación externa con datos provenientes de otros hospitales o bases de datos, lo que dificulta la generabilidad de los modelos utilizados.

2. **Heterogeneidad metodológica**: en la segmentación, la selección de características y la evaluación de modelos. Algunos trabajos emplean segmentaciones manuales o semiautomáticas realizadas por uno o pocos radiólogos, mientras que otros utilizan delineaciones automáticas basadas en redes convolucionales para cuantificar volúmenes y razones derivadas (p. ej., edema, necrosis o componente sólido), introduciendo variabilidad en la definición de las regiones de interés. De forma análoga, la selección de variables difiere sustancialmente entre estudios, abarcando desde cribas multietapa con filtrado de redundancia y selección supervisada hasta reducciones basadas en componentes principales, así como criterios empíricos de estabilidad en remuestreos repetidos; a ello se suma la coexistencia de esquemas de validación y métricas no uniformes, lo que dificulta la comparación directa y la reproducibilidad entre trabajos [2, 44, 29].

   De la misma forma, la selección de características difiere entre estudios: desde métodos clásicos de reducción de dimensionalidad (LASSO, RFE, PCA) hasta enfoques híbridos que combinan *hand-crafted features* y *deep transfer learning* [2, 22]. Esta disparidad metodológica se extiende también a la fase de validación, donde coexisten esquemas de partición muy heterogéneos (LOOCV, k-fold y validaciones internas repetidas), una escasa utilización de cohortes externas y una falta de estandarización en las métricas reportadas —predominando medidas de discriminación frente a análisis de calibración o utilidad clínica—, lo que dificulta la comparación cuantitativa entre estudios y limita la reproducibilidad de los resultados [5, 45].

3. **Falta de estandarización** En cuanto a la adquisición y secuencias en MRI, existen configuraciones muy dispares: Mientras que Solar se centra exclusivamente en DWI/ADC con análisis lineal de intensidades [38], Priya emplea un protocolo multiparamétrico (T1W, T2W, T1-CE, ADC, FLAIR; escáner 1.5 T) [30], Xiao combina CE-T1WI y T2-FLAIR y añade la razón edema/volumen tumoral [44], Bo utiliza T1WI/T2WI con enfoque híbrido entre *hand-crafted* + *deep transfer learning* [2] y Xia se apoya únicamente en CE-T1WI con *LightGBM* y explicación con SHAP [43].

   En cuanto a la segmentación y máscaras utilizadas también encontramos heterogeneidad: algunos trabajos extraen características del volumen tumoral completo y/o del edema [30, 44], mientras que otros comparan máscaras diferenciadas [30], y Pomohaci introduce segmentación automática con CNN frente a delineaciones manuales habituales [29].

   En relación al espacio de características y modelos usados coexisten radiómica

clásica (LASSO, LR/RF) [30, 44], enfoques híbridos HCR+DTL [2] y DL con segmentación/estimación integradas [29, 24], con métricas no uniformes (AUC, exactitud, etc.).

4. **Interpretabilidad clínica limitada**: Aunque algunos estudios recientes han comenzado a incorporar técnicas de interpretación como la estimación explícita de parámetros biológicos de crecimiento e infiltración tumoral a partir de DWI/ADC mediante modelos físico-informados [24], o el uso de SHAP para cuantificar la importancia relativa de las variables radiómicas [43], la mayoría de enfoques de alto rendimiento siguen funcionando como cajas negras. En trabajos clásicos de radiómica supervisada o de *deep transfer learning* [2, 22] el proceso de decisión no se vincula de forma explícita a hallazgos fisiopatológicos observables por el radiólogo. Existen excepciones puntuales con cierta trazabilidad clínica, como el uso de la razón edema/volumen tumoral para distinguir absceso de glioblastoma [44], la interpretación del gradiente de valores ADC en el anillo lesional frente al edema circundante [38]. Sin embargo, este tipo de interpretabilidad basada en rasgos anatómico-biológicos sigue sin ser la norma y no está sistematizada.

Observamos, como ya hemos mencionado, que la tomografía computarizada ha recibido escasa atención en la literatura radiómica sobre lesiones cerebrales infecciosas y tumorales. Además, la mayoría de literatura existente se ha centrado en la diferenciación entre el glioblastoma y tumores primarios o metástasis, sin abordar de forma sistemática el diagnóstico diferencial frente al absceso.

Por tanto, se identifica un vacío de conocimiento en la aplicación de la radiómica explicable sobre TC para distinguir entre GBM y AC.

2.0.3 Análisis cuantitativo comparativo

A continuación en la Figura 2.1 se muestra una síntesis numérica de las métricas de rendimiento reportadas en los estudios, tomando como referencia el AUC y, en su ausencia, el *accuracy* reportado. Para las revisiones, se tomó el valor medio de las métricas reportadas.

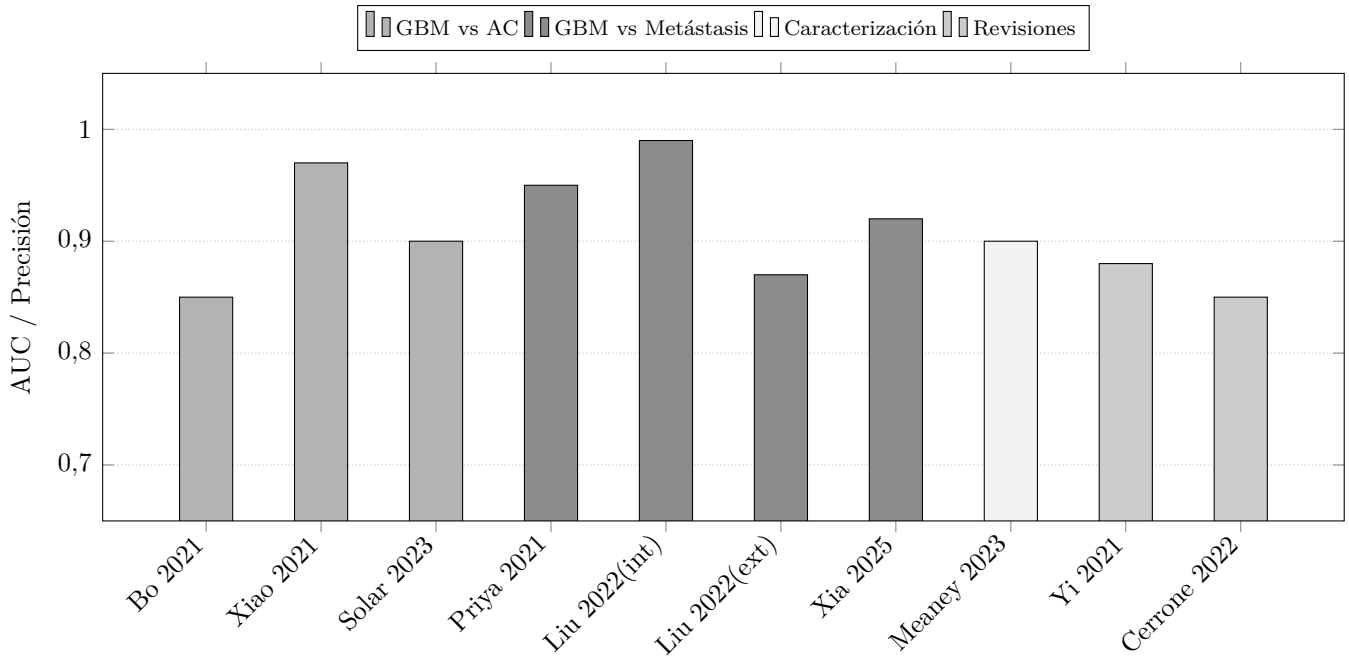


Figura 2.1: Comparación del AUC / Acc reportada por los principales estudios revisados. Cada barra representa el valor máximo de la métrica reportada en cada estudio. Para las revisiones, se tomó el valor medio de las métricas reportadas

En conjunto, los valores de rendimiento obtenidos muestran una alta capacidad discriminativa de los modelos evaluados. El AUC reportado entre los distintos estudios oscila entre un mínimo de aproximadamente 0.85 y un máximo de 0.99, con una media global en torno a 0.91. Específicamente, el AUC promedio de los trabajos que discriminan entre GBM y AC es de **0.9**. Estos valores reflejan una tendencia general a resultados cercanos a la excelencia diagnóstica en resonancia magnética, especialmente en los trabajos que emplean enfoques multiparamétricos o validación externa, mientras que las revisiones sistemáticas y los análisis de caracterización mantienen una precisión media ligeramente inferior (AUC $\approx$ 0.86–0.88).

2.0.4 Reducción de dimensionalidad y proyección supervisada

Los enfoques modernos de clasificación basados en radiómica operan con descriptores de muy alta dimensionalidad (forma, intensidad, texturas multiescala, gradientes locales, etc.), lo que introduce dos problemas clásicos: riesgo de sobreajuste en escenarios de muestra pequeña y dificultad para interpretar qué separa realmente una clase de otra [1]. Para mitigar esto, la literatura recurre a dos familias de estrategias: **selección de características** y **extracción/proyección de características**

En la selección de características, el objetivo es conservar un subconjunto de variables originales consideradas más relevantes (por ejemplo, mediante LASSO o *Recursive Feature Elimination*), de modo que el clasificador trabaje con menos predictores manteniendo interpretabilidad clínica. No obstante, en cohortes clínicas pequeñas, eliminar variables de forma agresiva puede descartar señal combinada distribuida en varias características débiles y, por tanto, degradar la capacidad discriminativa global. Este efecto es especialmente problemático en radiómica, donde el número de pacientes es limitado frente al número de *features*.

En cambio, las técnicas de extracción/proyección no seleccionan variables preexistentes, sino que generan nuevas variables en un espacio latente de menor dimensión que intenta concentrar la información relevante para la tarea. Entre las más empleadas se encuentran:

- **PCA (Análisis de Componentes Principales).** PCA es una proyección lineal no supervisada que construye combinaciones ortogonales de las variables originales maximizando la varianza explicada por los primeros componentes [18]. En imagen médica se usa de forma rutinaria para reducir colinearidad y ruido antes del entrenamiento de un clasificador. Sin embargo, presenta dos limitaciones claras en diagnóstico diferencial: (i) no utiliza etiquetas de clase (p.ej., absceso vs glioblastoma), de modo que preserva la estructura global de los datos pero no garantiza máxima separabilidad entre clases, y (ii) los componentes son combinaciones lineales potencialmente densas de cientos de *features*, lo que dificulta su trazabilidad clínica.

- **LDA (Linear Discriminant Analysis).** LDA es un método supervisado lineal clásico que busca una proyección donde las clases queden lo más separadas posible maximizando la varianza inter-clase y minimizando la intra-clase [37]. Su uso como paso previo a la clasificación ha sido históricamente eficaz cuando el número de predictores es moderado. Sin embargo, en dominios biomédicos de alta dimensionalidad y bajo número de pacientes (caso típico de radiómica cerebral), LDA se enfrenta al llamado *small sample size problem*: las matrices de dispersión intra-clase se vuelven mal condicionadas o singulares cuando el número de variables es comparable o superior al número de sujetos. Esto obliga a regularización adicional o a variantes modificadas, lo que limita su aplicabilidad directa.

- **Proyecciones que preservan estructura local (LPP, *Locality Preserving Projections*).** LPP formula la reducción de dimensionalidad como un problema de preservación de la geometría local: construye un grafo de vecindad entre las muestras en el espacio original y aprende una proyección lineal de baja dimensión que mantiene, en lo posible, esas relaciones de proximidad [16]. Este tipo de aproximación es especialmente atractivo en contextos donde las diferencias entre lesiones son locales y sutiles —p.ej., bordes irregulares, edema perilesional, cavitación necrótica— más que globales. Su desventaja en su formulación original es que se trata de un método esencialmente no supervisado: se preserva la geometría intrínseca de los datos, pero no se fuerza explícitamente la separación entre clases.

- **LOL (Linear Optimal Low-rank projection).** LOL es una técnica de proyección supervisada diseñada específicamente para escenarios de ultra-alta dimensionalidad y pocas muestras. Incorpora explícitamente momentos condicionales por clase (medias y covarianzas), combinando la intuición de LDA con la capacidad de escalar como PCA, y demostrando garantías teóricas de optimalidad y eficiencia computacional en contextos biomédicos con millones de *features* [42]. LOL ha mostrado superar a PCA, PLS y variantes regularizadas de LDA en tareas clínicas reales de neuroimagen y genómica con p mayor que n.

- **SNMF (Supervised Non-negative Matrix Factorization).** SNMF introduce supervisión sobre la factorización no negativa clásica incorporando directamente la función de pérdida del clasificador en el proceso de factorización [6]. Mientras que NMF estándar descompone la matriz de datos en componentes aditivos e

interpretables, SNMF además fuerza a que la representación latente preserve información discriminativa asociada a las etiquetas. En la práctica, esto se traduce en factores latentes más interpretables clínicamente (al ser no negativos y aditivos) y, simultáneamente, orientados a maximizar la capacidad predictiva. Este enfoque ha demostrado mejorar la estratificación pronóstica en entornos clínicos complejos, como la predicción de mortalidad en UCI, frente a variantes NMF puramente no supervisadas.

En los últimos años se han propuesto extensiones supervisadas de estas técnicas de proyección para abordar simultáneamente dos objetivos: mantener la estructura local relevante y, a la vez, maximizar la discriminación entre etiquetas. Entre ellas se encuentra **SLMVP (Supervised Local Maximum Variance Preserving)**, que introduce la información de clase en la construcción del grafo y en la función objetivo de la proyección [15]. En términos generales, SLMVP aprende una transformación que (i) preserva la estructura local de vecindad considerada clínicamente homogénea (por ejemplo, regiones radiológicamente similares dentro de una misma categoría diagnóstica) y (ii) amplifica las diferencias sistemáticas entre clases en el subespacio proyectado. Este planteamiento puede verse como una hibridación entre la filosofía geométrica de LPP y el criterio discriminante de LDA, específicamente adaptada al régimen de alta dimensionalidad y pocas muestras.

Para el problema diferencial entre glioblastoma y absceso cerebral, esta proyección supervisada tiene dos consecuencias prácticas directas:

1. **Mejora de la separabilidad en cohortes pequeñas.** Al incorporar supervisión explícita, la proyección prioriza las diferencias estructurales relevantes entre ambas entidades (p.ej., heterogeneidad infiltrativa y necrosis multilobulada en glioblastoma frente a cavidad central purulenta encapsulada en el absceso), incluso cuando dichas diferencias están dispersas en muchas *features* débiles si se evalúan de forma aislada.

2. **Facilitación de clasificadores estables post-proyección.** Tras reducir el espacio radiómico original a unas pocas dimensiones discriminativas, clasificadores relativamente simples y robustos al sobreajuste, como una SVM lineal, pueden operar con mejor estabilidad estadística y menor varianza en validación cruzada, lo que es crítico cuando el número total de pacientes es bajo.

En este trabajo se adopta precisamente este segundo enfoque: se aplica una etapa de proyección supervisada de tipo SLMVP previa al entrenamiento de una SVM binaria, con el objetivo de concentrar la información discriminativa entre absceso cerebral y glioblastoma en un subespacio compacto y más robusto. Esta estrategia se aleja de la práctica más común en la literatura clínica y constituye una de las principales aportaciones metodológicas de este TFM.

Capítulo 3

Material y métodos

3.0.1 Base de datos

Uno de los objetivos de este trabajo era el de proporcionar un número de muestras suficientemente amplio como para entrenar los modelos de clasificación. Con este fin, se incluyeron 62 muestras (31 glioblastomas y 31 abscesos) seleccionadas según los criterios establecidos por un radiólogo del Hospital Universitario Ramón y Cajal con más de veinte años de experiencia:

Características de interés:

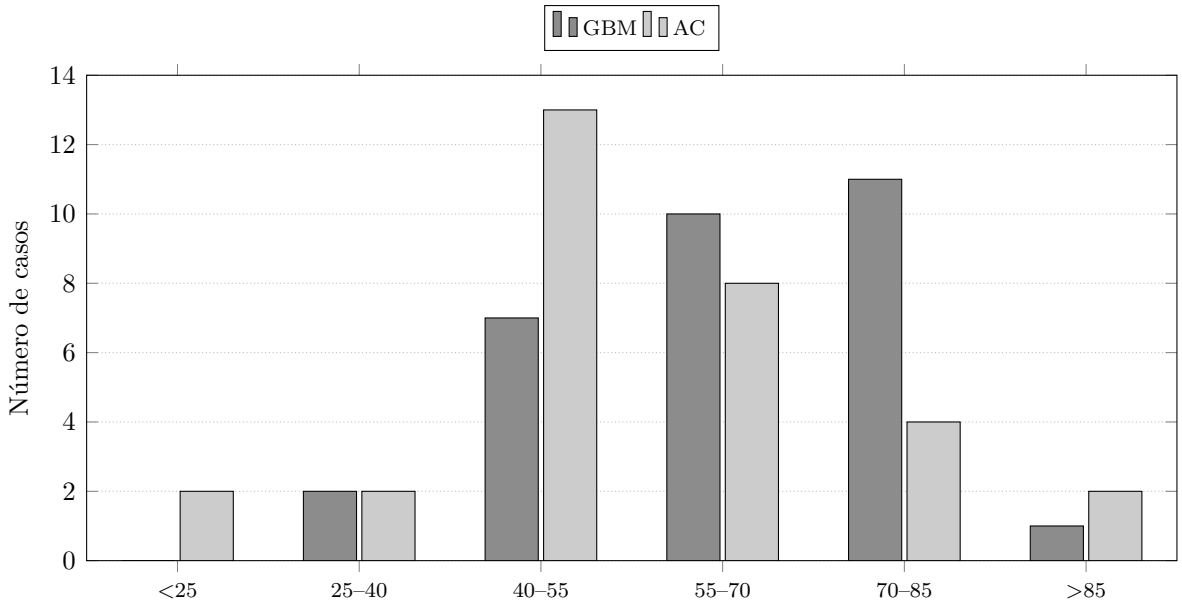
- Lesión con captación en anillo (realce periférico, liso o grosero, centro de baja densidad, necrótico, quístico...)

Se han excluido:

- Las lesiones sin realce en anillo evidente, lesiones con comportamiento sólido con captación prácticamente de toda la lesión, lesiones sangradas, lesiones no captantes...
- Las lesiones con diagnóstico erróneo (metástasis, hematoma, glioma de bajo grado)
- Las lesiones que tenían un estudio incompleto de TC y no disponían de CIV
- Las lesiones con estudio de TC postquirúrgico o postbiopsia.

Estas lesiones están pareadas por fecha de adquisición de la imagen (para evitar posibles cambios en el protocolo de adquisición), edad y sexo en ese orden cuando fue factible, debido a que la muestra de abscesos era limitada.

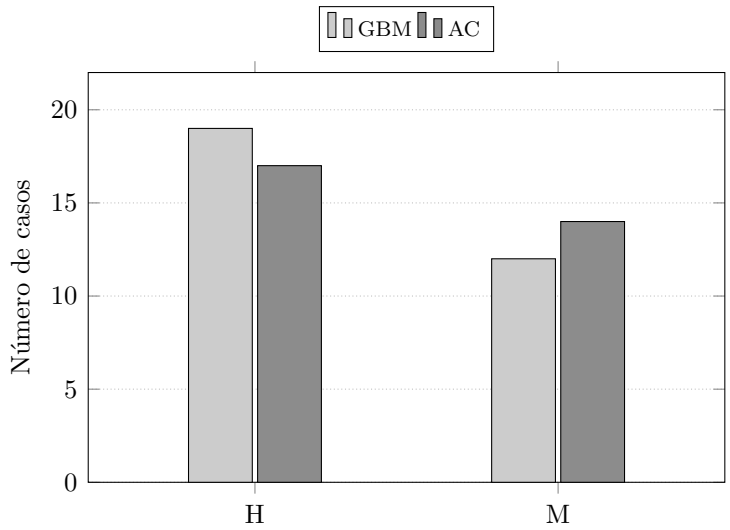
La distribución por edad de ambas lesiones se encuentra acotada en un rango de entre 18 y 91 años. Se observa que el grupo con mayor incidencia de GBM es el que se encuentra entre los 70 y los 85 años, mientras que los abscesos son más frecuentes en pacientes más jóvenes, entre los 40 y 55 años. Lo cual es coherente con la literatura [28] [25].



Los rangos de edad para cada lesión son:

- **Glioblastomas**: Promedio de 63,90 con un rango entre 39 y 88 años.
- **Abscesos**: Promedio de 55,74 con un rango entre 18 y 91 años.

En ambas patologías encontramos mayor incidencia masculina siendo un 58 % superior la incidencia de casos masculinos en GBM y un 21 % superior en AC. En el caso de los glioblastomas, esta tendencia se puede atribuir a factores biológicos y genéticos que predisponen a los hombres a desarrollar este tipo de tumor. En los abscesos cerebrales, esta mayor incidencia puede estar relacionada con factores predisponentes como la mayor incidencia de infecciones en hombres, y más en concreto aquellas que atacan el sistema inmune [25].



Las imágenes correspondientes a esta base de datos se encontraban almacenadas en el sistema PACS del Hospital Universitario Ramón y Cajal, un sistema informático que

permite capturar, almacenar, distribuir y visualizar imágenes médicas de manera digital, como radiografías, tomografías y resonancias magnéticas [14]. En el servicio de Urgencias del Hospital Universitario Ramón y Cajal, para la mayoría de las imágenes del estudio, se utilizó un escáner Toshiba Aquilon 64, un tomógrafo computarizado multicorte desarrollado por Toshiba (actualmente Canon Medical Systems) que permite obtener imágenes de alta resolución con 64 cortes simultáneos de 0,5 mm en cada rotación [26].

El protocolo de adquisición de imágenes fue homogéneo y estandarizado para todos los pacientes, garantizando consistencia en la calidad y comparabilidad de los estudios.

Aunque la mayoría de las imágenes provenían del sistema de tomografía computarizada de Toshiba del servicio de urgencias, dos de las imágenes de los abscesos cerebrales provenían de otros equipos diferentes, más concretamente Philips y Siemens. Esto se debe a las limitaciones en la obtención de la muestra inicial de abscesos, por lo que se decidió usar esas muestras pese a las posibles inhomogeneidades causadas por estas máquinas.

Con respecto al grosor de corte, se realizó un interpolado a 1 mm de grosor partiendo de cortes de 4 mm, ya que la mayor parte de la base de datos disponía de imágenes con este grosor de corte. Si bien es cierto que reconstruir a partir de cortes de 0,5 mm es más sencillo e incluye menos ruido en la imagen, tan sólo los estudios a partir del año 2015 comenzaban a tener esta resolución de corte y, por tanto, de nuevo perderíamos imágenes de valor para el entrenamiento del algoritmo de diferenciación si tan sólo se considerasen esas. Además, al realizarse el mismo tipo de interpolación en la totalidad de las imágenes de la base de datos, el ruido que esto pueda introducir es simétrico y, por tanto, no se tendrá en cuenta como elemento diferenciador de las lesiones cuando se extraigan las características de radiómica.

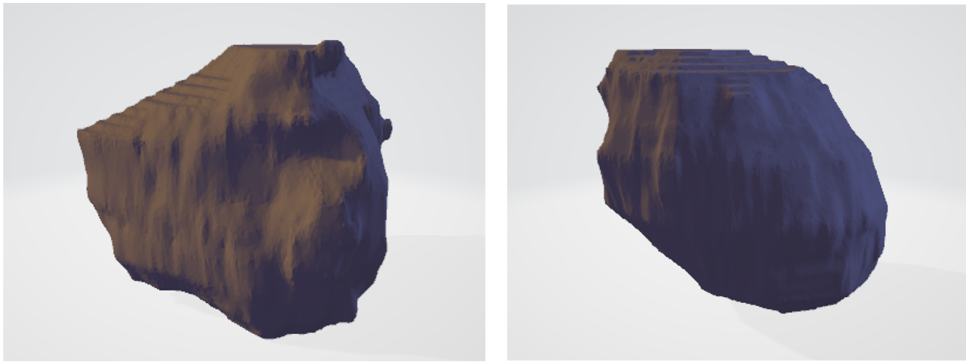


Figura 3.1: Visualización tridimensional de un glioblastoma (izquierda) y un absceso (derecha) segmentados usando Synapse3D.

La herramienta usada para realizar esta interpolación fue Synapse3D un software avanzado de visualización y procesamiento de imágenes médicas desarrollado por Fujifilm, que permite realizar reconstrucciones tridimensionales, análisis detallados y manipulaciones precisas de datos de imágenes médicas, optimizando el diagnóstico y la planificación de tratamientos [13].

3.0.2 Extracción de características

Para la extracción de las características se optó por utilizar el software de código abierto para el análisis y visualización de datos médicos *3D Slicer*, donde se seleccionaron todas las

características que pudieran resultar de interés [12], siendo un total de 107. Estas fueron:

1. **Características de forma:** Incluyen medidas como elongación, planitud, esfericidad, que describen la geometría de la lesión. Suman en total 14 características.

2. **Características de primer orden:** Relacionadas con la distribución de intensidades en la imagen, como la media, mediana, uniformidad, varianza, entre otras. Suman en total 18 características.

3. **GLCM (Gray Level Co-occurrence Matrix):** Mide la frecuencia con la que aparecen combinaciones específicas de niveles de gris en una imagen, capturando la textura basada en la co-ocurrencia de intensidades. Suman en total 24 características.

4. **GLDM (Gray Level Dependence Matrix):** Evalúa la dependencia de los niveles de gris en una imagen, midiendo cómo varía la intensidad de un píxel respecto a sus vecinos cercanos. Suman en total 14 características.

5. **GLRLM (Gray Level Run Length Matrix):** Mide la longitud de secuencias consecutivas de píxeles con el mismo nivel de gris, capturando patrones lineales en la textura. Suman en total 16 características.

6. **GLSZM (Gray Level Size Zone Matrix):** Analiza zonas homogéneas de niveles de gris en la imagen, midiendo el tamaño y la frecuencia de estas áreas. Suman en total 16 características.

7. **NGTDM (Neighboring Gray Tone Difference Matrix):** Evalúa las diferencias de intensidad entre un píxel y sus vecinos, capturando la variabilidad local en la textura. Suman en total 5 características.

No se aplicaron estrategias de reducción de dimensionalidad, ya bien basadas en correlación entre características, en métodos estadísticos como *Principal Component Analysis* (PCA) o *Independent Component Analysis* (ICA), o en técnicas de selección de características supervisadas como *Recursive Feature Elimination* (RFE). La razón de esto es que, aunque su empleo puede reducir la redundancia, también incrementa el riesgo de pérdida de información y la interpretabilidad del modelo. Se reconocieron varias limitaciones:

- **Pérdida potencial de información:** Características altamente correlacionadas no son necesariamente redundantes, y algunas pueden aportar matices relevantes para la clasificación.

- **Impacto en explicabilidad:** La eliminación de ciertas características más interpretables en favor de otras menos comprensibles puede dificultar la interpretación clínica o práctica de los resultados.

En contraste, la técnica SLMVP-SVM proyecta los datos en un subespacio de menor dimensión optimizando la separación entre clases e incorporando la información de etiquetas durante la proyección. De esta forma, SLMVP preserva o mejora la separabilidad de las categorías sin necesidad de eliminar previamente características.

Por lo tanto, a nivel teórico, la reducción manual de características mediante filtrado por

correlación no es necesaria en este trabajo. De esta forma tenemos varias ventajas:

- **Preservación de información útil:** Evita la eliminación arbitraria de variables que, aunque correladas, podrían contener información discriminativa complementaria.
- **Incorporación de información supervisada:** La proyección toma en cuenta las etiquetas de clase, priorizando la separación de categorías relevantes en lugar de simplemente reducir redundancia.
- **Robustez frente a alta dimensionalidad:** Los métodos basados en SVM y proyecciones lineales supervisadas pueden manejar un número elevado de características sin sufrir sobreajuste, siempre que se utilicen técnicas adecuadas de regularización.

3.0.3 Clasificación

El diseño experimental consta de dos estrategias metodológicas con el objetivo de optimizar la validación y la evaluación del rendimiento del modelo:

- **Validación cruzada exhaustiva:** En primer lugar, el conjunto completo se dividió en dos particiones independientes: un conjunto de entrenamiento (70 %) y un conjunto de prueba independiente (30 %).

  Dado el número limitado de muestras (62 casos en total, 31 por clase), dentro del conjunto de entrenamiento se empleó una validación cruzada estratificada con el máximo número de particiones posible, garantizando en cada *fold* la presencia de al menos una muestra de cada clase (absceso y glioblastoma). En la práctica, esto condujo a una validación cruzada con 21 *folds*, lo que equivale a una estrategia tipo *leave-one-out* sobre el conjunto de entrenamiento, ya que dividimos el conjunto de train en el máximo número de splits posibles.

  Este esquema permite que cada muestra del conjunto de entrenamiento actúe sucesivamente como validación interna mientras el resto se utiliza para el ajuste del modelo, maximizando el aprovechamiento de la información disponible y reduciendo la varianza asociada a particiones arbitrarias de los datos.

  Como excepción, en el caso del modelo *Explainable Boosting Machine (EBM)*, el número de *folds* se redujo a 8 debido a su mayor coste computacional, manteniendo en todo momento la estratificación y el equilibrio entre ambas clases.

- **Evaluación final externa:** Una vez optimizados los hiperparámetros mediante la validación cruzada interna, el modelo final se entrena nuevamente sobre todo el conjunto de entrenamiento y se evalúa sobre el conjunto de prueba independiente. Sobre este conjunto se calculan las principales métricas de rendimiento: *accuracy*, *F1-score*, *recall*, *precision* y área bajo la curva ROC (*AUC*). Estas métricas se obtienen a partir de las predicciones discretas y de las probabilidades asociadas a la clase positiva (glioblastoma), permitiendo evaluar tanto la capacidad global de clasificación como el equilibrio entre sensibilidad y especificidad. El uso de un conjunto de test completamente separado garantiza una estimación imparcial del desempeño real del modelo en datos no vistos durante el entrenamiento, lo cual resulta de especial interés para su aplicación en la práctica clínica.

Para obtener una estimación más robusta del rendimiento, el procedimiento completo se repite un total de **15 veces** con distintas particiones aleatorias train-test manteniendo siempre la proporción 70-30. Esto es fundamental ya que el rendimiento del modelo depende mucho del conjunto de test utilizado debido precisamente al tamaño del dataset, encontrando tan sólo 19 muestras en el conjunto de test.

Las métricas finales reportadas corresponden al promedio y desviación estándar de los valores obtenidos en estas repeticiones, lo que permite mitigar la influencia de la variabilidad muestral y de los posibles efectos de una única partición particular. La desviación estándar da cuenta de la estabilidad de cada métrica reportada.

Este enfoque metodológico se aplicó usando cuatro modelos distintos de *machine learning*: Random Forest, XGBoost, EBM y SVM, con el objetivo de evaluar tanto algoritmos basados en conjuntos de árboles como métodos lineales y modelos aditivos explicables, y así disponer de una visión amplia del rendimiento alcanzable bajo distintos supuestos inductivos (no linealidad, interacciones, interpretabilidad, etc.). Finalmente, se aplica el modelo con proyección de características SLMVP–SVM para comparar sus resultados con los anteriores y determinar hasta qué punto la incorporación de una proyección supervisada orientada a maximizar la separabilidad local mejora la estabilidad estadística, la capacidad discriminativa y la robustez del clasificador frente al reducido tamaño muestral.

Modelo 1: Random Forest

Para el primer modelo se empleó un clasificador **Random Forest**. Random Forest [3] construye un *ensamble* de árboles de decisión independientes mediante muestreo con reemplazo del conjunto de entrenamiento (*bagging*) y selección aleatoria de subconjuntos de características en cada división del árbol. Esto introduce variabilidad estructural entre los árboles, reduciendo su correlación y, por tanto, la varianza global del modelo. Cada árbol aprende un mapeo no lineal entre las características y la clase, y la predicción final se obtiene mediante una votación mayoritaria entre todas las predicciones individuales. Breiman demostró que el error de generalización del bosque está acotado por

$$\mathrm{Err}_{RF} \leq \frac{\rho\,(1-s^2)}{s^2},$$

donde s representa la *fuerza* media de los árboles (capacidad discriminativa individual) y ρ la correlación entre ellos. Así, el rendimiento mejora al incrementar la independencia entre árboles y al mantener árboles suficientemente expresivos, lo cual se consigue precisamente mediante la combinación de *bootstrap* y selección aleatoria de características.

Se evaluaron múltiples configuraciones del modelo variando el número de árboles, la profundidad máxima, el criterio de selección de características en los nodos y los parámetros de regularización (`min_samples_split`, `min_samples_leaf`), eligiéndose finalmente la combinación que maximizaba el AUC medio en la validación cruzada. Con los mejores hiperparámetros obtenidos, el modelo se reentrenó íntegramente sobre el *train* y posteriormente se evaluó sobre el subconjunto de prueba independiente.

En este estudio, el modelo resultante combina cientos de árboles de baja a moderada profundidad, ajustados para equilibrar capacidad y robustez, aplicando además pesos de clase. Su arquitectura hace que Random Forest sea especialmente adecuado en escenarios con un número elevado de variables predictoras y relaciones no lineales, proporcionando una reducción significativa de la varianza y un comportamiento estable frente al ruido. Esta

configuración permitió capturar patrones complejos en los datos y obtener un rendimiento reproducible a lo largo de todas las iteraciones de entrenamiento–evaluación.

Modelo 2: Explainable Boosting Machine (EBM)

Para el segundo modelo se empleó una **Explainable Boosting Machine (EBM)**, un clasificador interpretable basado en *gradient boosting* desarrollado por Microsoft Research. La EBM combina la estructura aditiva y transparente de los **Modelos Aditivos Generalizados (GAM)** con técnicas modernas de bagging y boosting, lo que permite capturar patrones complejos manteniendo una interpretación directa de la contribución de cada característica [33]. El modelo aprende cada función unidimensional $f_j(x_j)$ mediante un procedimiento *round-robin* con un *learning rate* bajo, reduciendo la colinealidad y evitando que el orden de las características influya en los resultados. Además, puede detectar interacciones de segundo orden $f_{i,j}(x_i, x_j)$ de manera automática cuando estas aportan mejoras visibles en el ajuste, manteniendo la interpretabilidad de la estructura global.

En la práctica, cada término aditivo se ajusta mediante pequeños árboles de regresión muy poco profundos (*shallow trees*), lo que proporciona curvas suaves y fácilmente interpretables incluso en contextos con pocas muestras. Esta arquitectura facilita visualizar el efecto marginal de cada variable y entender cómo contribuye a la predicción final, algo especialmente valioso en escenarios clínicos donde la transparencia del modelo es un requisito crítico.

Dado el elevado coste computacional asociado al ajuste de modelos EBM en un esquema de validación cruzada repetida, se optó por limitar el número de particiones internas de la validación cruzada a 8 *folds*. Esta decisión permite reducir de forma significativa el tiempo de entrenamiento sin comprometer la estabilidad de la estimación del rendimiento, manteniendo un equilibrio razonable entre robustez estadística y viabilidad computacional en un escenario de baja muestra.

Los hiperparámetros óptimos seleccionados tras la búsqueda incluyeron valores como `learning_rate`, `max_leaves`, `min_samples_leaf` y `max_bins`, que controlan la suavidad de las funciones aprendidas, la complejidad de los árboles base y la granularidad de las discretizaciones internas. En conjunto, la EBM proporcionó un equilibrio sólido entre interpretabilidad y capacidad predictiva, capturando relaciones no lineales relevantes sin perder trazabilidad en la explicación del modelo.

Modelo 3: XGBoost

En el tercer modelo se empleó *XGBoost* (Extreme Gradient Boosting), un algoritmo de *boosting* de árboles de decisión que optimiza de manera secuencial un modelo aditivo de árboles poco profundos, mejorando en cada iteración los errores residuales del conjunto anterior [7]. A diferencia de otros enfoques clásicos de *gradient boosting*, XGBoost introduce una formulación explícita de regularización y una implementación altamente optimizada que incluye paralelización a nivel de nodo, manejo eficiente de datos dispersos y soporte para grandes volúmenes de datos.

El algoritmo minimiza una función objetivo que combina la pérdida de entrenamiento y un término de complejidad del modelo. En la iteración t, la contribución del nuevo árbol f_t se aproxima mediante una expansión de segundo orden de la función de pérdida alrededor de

las predicciones actuales, de forma que el objetivo a optimizar queda dado por

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t),$$

donde g_i y h_i son el gradiente y el *hessiano* de la pérdida respecto a la predicción del modelo en la muestra i, y $\Omega(f_t)$ es un término de regularización que penaliza la complejidad del árbol, típicamente definido como

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2,$$

siendo T el número de hojas, w_j el valor de predicción de cada hoja, γ el coste de añadir una hoja adicional y λ un parámetro de regularización L_2 sobre los pesos. Esta formulación permite que XGBoost controle de manera explícita el sobreajuste, favoreciendo modelos que sean a la vez expresivos y parsimoniosos.

En este estudio, el modelo base de XGBoost se configuró como un clasificador de árboles gradualmente potenciados, en el que se ajustaron los principales hiperparámetros estructurales y de regularización. En particular, se exploraron distintas profundidades máximas de los árboles (`max_depth`), tasas de aprendizaje (`learning_rate`) y número de estimadores (`n_estimators`), así como parámetros de muestreo estocástico de instancias (`subsample`) y características (`colsample_bytree`), junto con términos de control de complejidad como `min_child_weight`, `gamma` y los coeficientes de regularización `reg_lambda` y `reg_alpha`. Esta combinación de *boosting* gradiente, regularización explícita y muestreo parcial de datos y variables hace que XGBoost resulte especialmente adecuado para conjuntos de datos biomédicos de alta dimensionalidad y potencialmente ruidosos, permitiendo capturar interacciones no lineales complejas entre las características y ofrecer un rendimiento estable a lo largo de las distintas particiones de entrenamiento y prueba.

Modelo 4: Support Vector Machine (SVM)

El cuarto modelo empleado fue un clasificador **Support Vector Machine (SVM)**, introducido por Cortes y Vapnik [9]. El principio fundamental de las SVM consiste en encontrar un hiperplano que maximice el *margen* entre clases, es decir, la distancia entre dicho hiperplano y los ejemplos más cercanos de cada categoría (los *vectores soporte*). Esta formulación geométrica proporciona una garantía teórica de robustez y un elevado poder de generalización, especialmente en contextos con muchas características y pocas muestras. Para datos no separables linealmente, el modelo introduce un término de penalización que controla el equilibrio entre maximizar el margen y permitir errores de clasificación, y para capturar relaciones no lineales incorpora funciones núcleo que proyectan los datos en espacios de mayor dimensionalidad sin necesidad de calcular explícitamente dicha transformación.

El modelo se evaluó con dos funciones de núcleo: *lineal* y *RBF* (*Radial Basis Function*). El núcleo RBF permite modelar relaciones altamente no lineales entre las variables, ajustando su flexibilidad mediante el parámetro γ, mientras que el núcleo lineal proporciona una hipótesis más simple y fácilmente interpretable, adecuada en escenarios donde la frontera de decisión es aproximadamente lineal.

El hiperparámetro principal del modelo, C, controla el grado de penalización asignada a los errores de clasificación: valores altos fuerzan una separación estricta entre clases,

mientras que valores reducidos favorecen márgenes más amplios y mayor regularización. En los modelos con núcleo RBF, el parámetro γ determina la influencia espacial de cada muestra, afectando directamente a la suavidad de la frontera de decisión.

En este trabajo, todas las características fueron estandarizadas previamente, un paso imprescindible en SVM debido a su sensibilidad a la escala de los predictores. El modelo resultante permitió establecer una referencia sólida para comparar posteriormente el clasificador SLMVP–SVM, analizando si la incorporación de una proyección supervisada previa es capaz de mejorar la separabilidad geométrica de las clases en contextos de baja muestra y alta dimensionalidad.

Modelo 5: Supervised Local Maximum Variance Preserving + SVM (SLMVP–SVM)

El quinto modelo integra un reductor de dimensionalidad **Supervised Local Maximum Variance Preserving (SLMVP)** con un clasificador **Support Vector Machine (SVM)**, combinando así una proyección supervisada no lineal con un modelo discriminativo robusto [15]. La idea fundamental de SLMVP es aprender una transformación B tal que los datos proyectados $P = B^\top X$ conserven simultáneamente: (i) la *estructura local* del espacio de entrada y (ii) la *variabilidad asociada a la clase* en el espacio de salida. Para ello se construyen dos grafos: uno basado en la similitud entre instancias en el espacio original y otro basado en la similitud entre etiquetas en el espacio de clases. Ambos grafos se expresan mediante matrices núcleo (K_x, K_y), las cuales se centran en el espacio de características para eliminar sesgos de traslación.

El objetivo de SLMVP consiste en maximizar la siguiente traza

$$\max_{B} \ \mathrm{tr}\left(B^\top X K_x K_y X^\top B\right),$$

de forma que se preserve la varianza local guiada por la información supervisada. La solución de este problema se obtiene mediante una descomposición en valores singulares en el espacio inducido por el núcleo, lo que permite generar un subespacio reducido de dimensión r (el *rank*). A diferencia de métodos clásicos como PCA o LDA, SLMVP no asume linealidad ni gaussianidad en los datos, pudiendo capturar deformaciones complejas del espacio mediante el núcleo RBF. Además, a diferencia de LDA, el número de dimensiones resultante no está limitado a $C - 1$, lo que otorga mayor flexibilidad en problemas de dos clases.

El parámetro clave del método es la **dimensión proyectada** (*rank* r), que determina cuántos componentes supervisados se conservan en el nuevo espacio. Valores de r bajos producen proyecciones muy compactas pero con riesgo de pérdida de información discriminativa, mientras que valores elevados pueden introducir ruido y redundancia. Por este motivo, se evaluó un conjunto amplio de ranks, desde $r = 1$ hasta un límite manejable impuesto por el número de muestras y la dimensionalidad de entrada, analizando cómo la capacidad predictiva varía con la complejidad de la proyección.

Tras la reducción supervisada, las características proyectadas alimentan una SVM, encargada de construir la frontera de decisión óptima en un espacio ya transformado para maximizar la separabilidad entre clases. Esta arquitectura ofrece varias ventajas clave:

- Mejora del margen separador al eliminar información no relevante y alinear el espacio con la estructura de clases;

- Mitigación del problema *"p grande y m pequeña"* al reducir drásticamente la dimensionalidad previa al entrenamiento;

- Preservación de vecindarios y variabilidad local, que se traduce en fronteras más coherentes en entornos con alta no linealidad.

El enfoque SLMVP–SVM puede interpretarse así como un esquema en dos etapas: primero, una proyección supervisada explícitamente diseñada para realzar la separabilidad geométrica; después, un clasificador de máximos márgenes con elevada capacidad de generalización. Los resultados obtenidos muestran que el rendimiento del modelo depende de forma notable de la elección del rank, existiendo un punto óptimo donde el subespacio conserva suficiente variabilidad discriminativa sin sobredimensionarse. Este comportamiento confirma que la combinación SLMVP–SVM explota eficazmente la información supervisada durante la reducción de dimensionalidad, proporcionando mejoras significativas frente al uso de SVM sin proyección previa en contextos de alta dimensionalidad y muestras limitadas.

3.0.4 Explicabilidad

La capacidad de explicar el funcionamiento de los modelos de aprendizaje automático resulta fundamental, en particular dentro del ámbito médico, donde las decisiones determinan de manera directa la atención y posterior tratamiento de los pacientes. Entender el proceso mediante el cual un modelo genera sus predicciones no solo aumenta la confianza en los resultados obtenidos, sino que también favorece su adopción y uso por parte de los profesionales de la salud.

En general se puede decir que cuanto más explicable es un modelo, menor es su capacidad predictiva, pues la búsqueda de interpretaciones claras suele implicar estructuras menos complejas. Sin embargo, en áreas críticas como la médica, es indispensable equilibrar la robustez de las predicciones con la transparencia sobre cómo se obtienen. De este modo, se gana la confianza de los profesionales de la salud y se facilita la adopción de estos modelos en la práctica clínica. Para lograrlo, se han desarrollado enfoques que, sin sacrificar demasiado rendimiento, brindan explicaciones detalladas sobre la contribución de cada variable en la toma de decisiones.

SHAP en modelos estándar

En el trabajo previo se empleó el método SHAP (Shapley Additive Explanations) [23], que se basa en la teoría de juegos para asignar a cada característica un valor de importancia específico en una predicción determinada. En SHAP, cada característica es considerada como un jugador en un juego cooperativo, donde el objetivo es predecir el resultado del modelo. Se generan múltiples juegos (subconjuntos de características) y luego se evalúa la contribución de cada jugador (característica) al resultado en todos los posibles subconjuntos.

Concretamente, el *valor Shapley* es la contribución media de una característica a lo largo de todas las combinaciones posibles de características. Matemáticamente, para una característica i, se define como:

$$\Phi_i(V) = \sum_{S \subseteq D \setminus \{i\}} \frac{|S|!\,(|D| - |S| - 1)!}{|D|!} \Big[v(S \cup \{i\}) - v(S) \Big]$$

donde:

- D es el conjunto de todas las características.
- S es una coalición sin la característica i, y $(S \cup \{i\})$ representa la contribución al añadir el jugador i.
- $v(S)$ es el valor del modelo con las características en S.

SHAP descompone la predicción de un modelo en contribuciones atribuibles a cada característica de entrada, ofreciendo varias ventajas:

1. **Transparencia:** Al desglosar la predicción en componentes individuales, permite a los profesionales médicos entender cómo cada característica influye en la decisión final, facilitando una interpretación más clara de los resultados.
2. **Consistencia:** SHAP garantiza que si una característica tiene mayor influencia en la predicción, esto se reflejará proporcionalmente en las explicaciones, proporcionando una base lógica y objetiva para la toma de decisiones clínicas.
3. **Visualizaciones efectivas:** Las herramientas de visualización de SHAP permiten representar gráficamente las contribuciones de las características, ayudando a identificar patrones y relaciones de manera intuitiva y accesible para los clínicos.

Este enfoque unificado para interpretar predicciones proporcionó una medida coherente y precisa de la contribución de cada característica al resultado del modelo, permitiendo compararlos posteriormente entre sí y extraer las características más relevantes que distinguen a las dos lesiones.

Métodos específicos de SHAP

En [23] se mencionan diferentes métodos específicos para calcular los valores SHAP según el tipo de modelo utilizado, tales como:

- **Tree SHAP:** Diseñado para modelos basados en árboles de decisión, permite calcular valores SHAP de forma exacta y eficiente aprovechando la estructura del árbol.
- **Kernel SHAP:** Método agnóstico al modelo. (*model-agnostic*) que puede aplicarse a cualquier tipo de modelo, utilizando un enfoque basado en muestras para aproximar los valores SHAP.
- **Deep SHAP:** Diseñado específicamente para redes neuronales profundas, combina SHAP con técnicas como DeepLIFT.
- **Linear SHAP:** Ideal para modelos lineales y de regresión logística, calcula las contribuciones lineales de manera eficiente.

En este trabajo, dado que se evaluaron modelos de distinta naturaleza, se adoptó una estrategia híbrida para el cálculo de valores SHAP en función de la familia del modelo. En concreto, se empleó **TreeSHAP** en los modelos basados en árboles (*Random Forest* y *XGBoost*) y **KernelSHAP** en los modelos no arbóreos (*EBM* y *SVM*). Esta decisión se fundamenta en que TreeSHAP explota la estructura interna de los árboles para calcular explicaciones de forma exacta y, sobre todo, mucho más eficiente computacionalmente en *ensembles* de árboles, siendo la opción natural para *Random Forest* y *XGBoost*. En cambio, *EBM* (modelo aditivo con boosting) y *SVM* (especialmente con kernels no lineales) no se ajustan a la formulación de árboles requerida por TreeSHAP, por lo que se recurrió a KernelSHAP, un enfoque *model-agnostic* que aproxima los valores Shapley mediante muestreo de coaliciones y evaluación del modelo como caja negra.

Al aplicar SHAP, se generaron dos visualizaciones principales para interpretar el comportamiento global del clasificador, ilustradas mediante ejemplos genéricos en las Fig. 3.2 y Fig. 3.3:

- **Gráfico *beeswarm*.** Resume simultáneamente *importancia* y *dirección del efecto* de cada característica. Cada fila corresponde a una característica y cada punto a una muestra. El eje horizontal representa el valor SHAP (contribución de esa característica respecto al valor base): valores positivos desplazan la salida del modelo hacia una mayor probabilidad de la clase objetivo, y valores negativos la reducen. La dispersión horizontal refleja la variabilidad del efecto entre individuos. Además, el color del punto codifica el *valor de la característica* (bajo–alto), lo que permite detectar patrones consistentes (p. ej., si valores altos de una variable empujan sistemáticamente la predicción en una dirección). Habitualmente, las características se ordenan de arriba a abajo por su contribución global (media de |SHAP|).

- **Gráfico de barras (*SHAP bar*) Proporciona una medida directa de importancia global** por característica, calculada como la media del valor absoluto de SHAP a lo largo de todas las muestras, $\mathbb{E}(|\Phi_i|)$. A diferencia del *beeswarm*, este gráfico no muestra la dirección (signo) del efecto, sino únicamente su magnitud promedio, facilitando un ranking claro de las variables más relevantes.

En conjunto, la combinación de TreeSHAP (para modelos arbóreos) y KernelSHAP (para modelos no arbóreos) permite mantener coherencia interpretativa entre clasificadores heterogéneos, optimizando el coste computacional cuando es posible y proporcionando explicaciones comparables para el análisis clínico posterior.

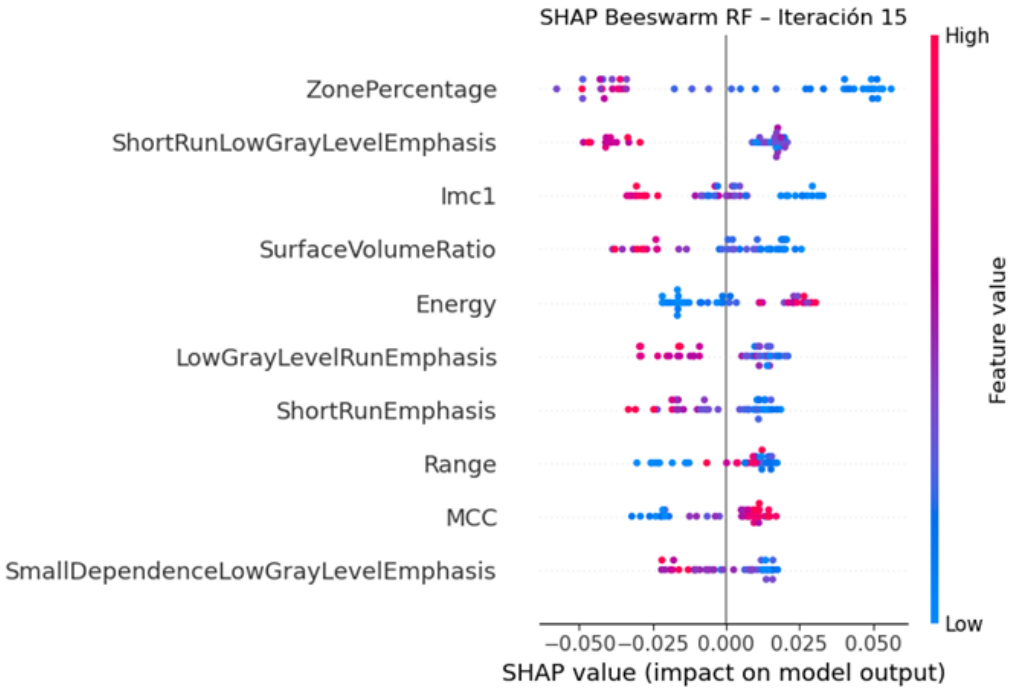


Figura 3.2: Ejemplo genérico de gráfico SHAP tipo *beeswarm* en Random Forest. Cada punto representa una muestra y su posición horizontal indica el valor SHAP. El color codifica el valor de la característica: tonos azules indican valores bajos o negativos y tonos rojos valores altos o positivos, permitiendo analizar cómo el valor de la variable se asocia con la dirección y magnitud de su efecto.

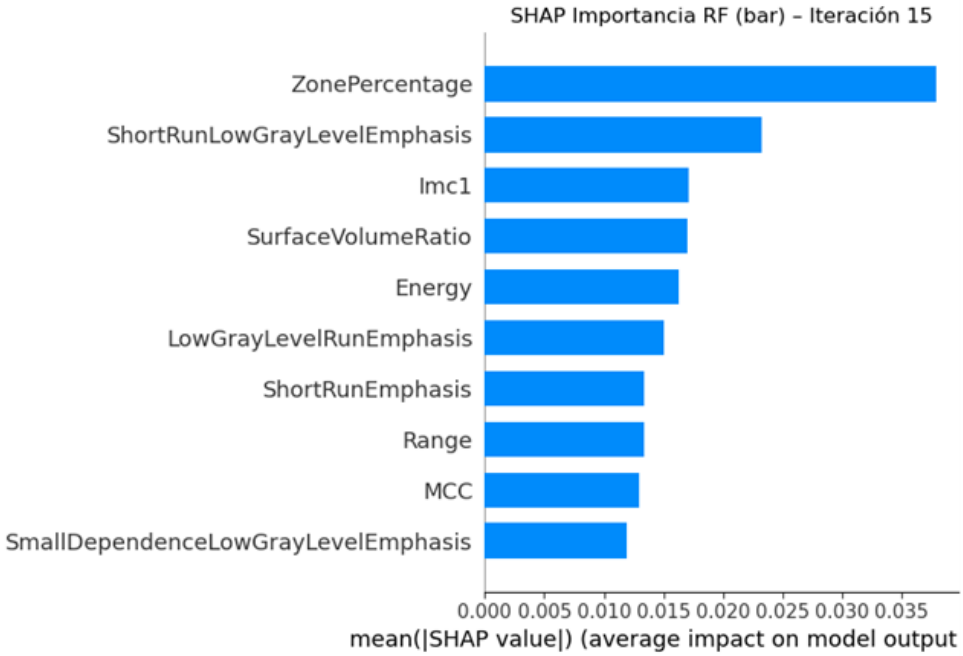


Figura 3.3: Ejemplo genérico de gráfico SHAP de barras en Random Forest, que muestra la importancia global de cada característica calculada como la media del valor absoluto de SHAP a lo largo de todas las muestras ($\mathbb{E}(|\Phi_i|)$).

SHAP en SLMVP-SVM

En el caso del modelo combinado SLMVP-SVM no es posible aplicar directamente ninguno de los métodos estándar de SHAP (TreeSHAP, DeepSHAP o LinearSHAP), ya que el clasificador final no es ni un árbol, ni una red neuronal profunda, ni un modelo estrictamente lineal sobre las variables originales, sino una SVM entrenada sobre una proyección supervisada previa (SLMVP) aprendida a partir de las propias etiquetas. En este trabajo se optó por una implementación propia de SHAP directamente sobre el *pipeline* completo:

De forma general, el funcional a explicar viene dado por

$$f(x) = \mathbb{P}\Big(\text{clase positiva} \mid x\Big),$$

donde la salida del modelo corresponde a la probabilidad estimada de glioblastoma para una muestra x.

Esta aproximación incluye explícitamente las tres etapas del *pipeline*: (1) estandarización de las características originales, (2) proyección SLMVP al subespacio de rango r, y (3) clasificación mediante SVM. En notación compacta, para una muestra $x \in \mathbb{R}^d$,

$$\tilde{x} = \frac{x - \mu}{\sigma}, \qquad z = B^\top \tilde{x} \in \mathbb{R}^r, \qquad f(x) = \hat{p}(y = 1 \mid z),$$

donde μ y σ son los parámetros del *StandardScaler*, B es la matriz de proyección SLMVP aprendida en cada iteración y $\hat{p}$ la probabilidad devuelta por la SVM tras su calibración interna.

Definición del funcional y espacio de explicación

El objetivo es asignar a cada característica original x_i un valor de Shapley $\phi_i(x)$ que mida la contribución marginal de dicha característica a la probabilidad $f(x)$ de pertenecer a la clase positiva. Se trabaja, por tanto, directamente en el espacio de características originales, pero evaluando siempre el funcional completo $f(\cdot)$ que incluye SLMVP+SVM, en línea con el enfoque agnóstico al modelo descrito en [46].

Dado un subconjunto de características $S \subseteq D = \{1, \ldots, d\}$ y una forma de "apagar" las restantes, se define el valor del modelo condicionado a la coalición S como

$$v_x(S) \;=\; \mathbb{E}\Big[f(x_S, X_{D\setminus S})\Big],$$

siguiendo la definición de contribución basada en expectativas condicionales introducida en [46]. El valor de Shapley de la característica i para la muestra x queda, como es habitual, dado por

$$\phi_i(x) \;=\; \sum_{S \subseteq D\setminus\{i\}} \frac{|S|!\,(|D| - |S| - 1)!}{|D|!}\Big[v_x(S \cup \{i\}) - v_x(S)\Big],$$

que garantiza propiedades deseables como eficiencia, simetría y nulidad de variables irrelevantes [46].

En la práctica, el cálculo exacto de esta expresión es intratable para dimensiones moderadas, por lo que se recurre a una aproximación Monte Carlo mediante permutaciones aleatorias de las características, tal y como se propone en [46].

Elección del conjunto de referencia (*background*)

Un aspecto crítico es establecer cómo se apagan las características que no forman parte de una coalición S. Se analizaron dos opciones:

1. **Baseline fijo global:** construir un único vector de referencia con valores medios o medianos. Este enfoque puede introducir combinaciones poco realistas de variables y romper correlaciones importantes, una limitación ya señalada en el marco general de explicaciones basadas en expectativas.

2. **Baselines reales muestreados:** definir un conjunto de fondo a partir de muestras reales del conjunto de entrenamiento, muestreadas según la distribución empírica de los datos, como sugiere el enfoque de muestreo de instancias y descrito en [46]. En cada permutación, las características apagadas se sustituyen por los valores de una muestra real, preservando mejor la estructura de covarianza.

Dado que el modelo SLMVP-SVM es especialmente sensible a las correlaciones entre variables de entrada, se optó por la segunda opción, ya que genera estados intermedios más plausibles y contribuciones marginales más interpretables.

Aproximación Monte Carlo de los valores de Shapley

Dado que el cálculo exacto de $\phi_i(x)$ es intratable para dimensiones moderadas, se adopta una aproximación Monte Carlo basada en permutaciones aleatorias de las características, siguiendo el **Algoritmo 1** propuesto por Štrumbelj y Kononenko [46]. Dicho algoritmo estima las contribuciones marginales evaluando el cambio en la salida del modelo al encender progresivamente las características según un orden aleatorio, manteniendo el resto fijado a un baseline (una instancia de referencia del conjunto de datos) muestreado aleatoriamente de la distribución empírica.

Sea $f(\cdot)$ el *pipeline* completo y x una muestra a explicar. El algoritmo implementado aproxima los valores de Shapley mediante M permutaciones aleatorias de las d características. Para cada muestra x:

1. Se inicializa un vector de contribuciones $\phi(x)$ a cero.

2. Para cada permutación:

   (a) Se selecciona de forma aleatoria un baseline real z.

   (b) Se genera una permutación π de las características.

   (c) Se parte del estado completamente apagado $x^{(0)} = z$ y se evalúa $f^{(0)} = f(x^{(0)})$.

   (d) Las características se encienden progresivamente según π, acumulando en cada paso la contribución marginal

   $$\Delta f^{(k)} = f^{(k)} - f^{(k-1)}$$

   en la característica correspondiente.

3. El valor final se obtiene como el promedio sobre las M permutaciones.

En este trabajo se utiliza $M = 100$. Para justificar esta elección, se evaluó la **convergencia**

de las importancias globales frente al número de permutaciones M. En concreto, para cada M se calcularon valores SHAP aproximados y se agregaron por característica como

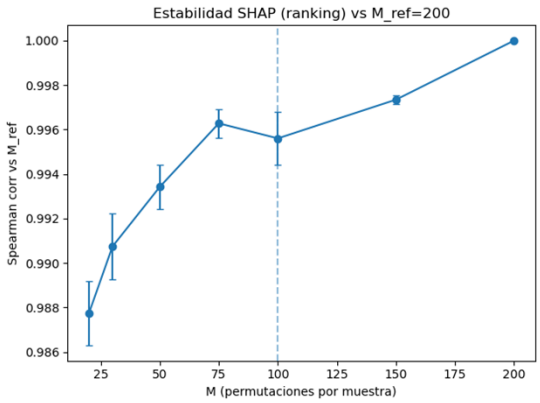
$$v_M = \frac{1}{n} \sum_{i=1}^{n} |\phi_M(x_i)| ,$$

comparándolos con una referencia $v_{M_{\text{ref}}}$ obtenida con un número mayor de permutaciones (M_{ref}). La estabilidad de las importancias se cuantificó mediante los siguientes criterios:

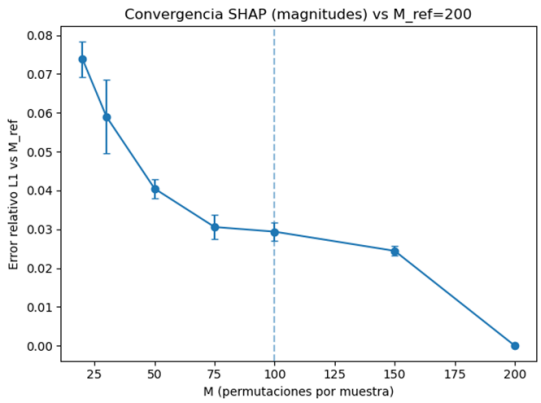
- **Correlación de Spearman** $\rho(v_M, v_{M_{\text{ref}}})$, como medida de consistencia de los valores SHAP en el *ranking* de características.
- **Error relativo** L_1 **normalizado**: mide cuánto difiere el vector de importancias globales (magnitudes absolutas) obtenido con M permutaciones respecto a una referencia con M_{ref}, normalizando por la norma L_1 de la referencia (error relativo):

$$\text{Err}_{L_1}^{\text{rel}}(M) = \frac{\|v_M - v_{M_{\text{ref}}}\|_1}{\|v_{M_{\text{ref}}}\|_1} = \frac{\sum_{j=1}^{d} |v_{M,j} - v_{M_{\text{ref}},j}|}{\sum_{j=1}^{d} |v_{M_{\text{ref}},j}|}.$$

Los resultados de cada uno de estos análisis se pueden ver en las Figuras 3.4a y 3.4b:



(a) Correlación de Spearman entre v_M y $v_{M_{\text{ref}}}$.



(b) Error relativo L_1 normalizado respecto a M_{ref}.

Figura 3.4: Análisis de convergencia de las importancias SHAP en función del número de permutaciones M. Se muestran la estabilidad de las magnitudes y la robustez del conjunto de variables más relevantes respecto a una referencia con M_{ref} permutaciones.

Uso de la probabilidad frente a la función de decisión

Aunque la SVM produce internamente una función de decisión $g(z)$, esta se transforma mediante una función sigmoide para obtener probabilidades. Trabajar con $g(z)$ o con la transformación logit de la probabilidad supone una transformación monótona del funcional explicado. Por razones de interpretabilidad clínica, se opta por explicar directamente la probabilidad $f(x)$ de la clase positiva (en lugar de los valores crudos proporcionados por la SVM), dado que así en [46] los valores de shap obtenidos también tienen una interpretación directa al usarse en modelos de regresión.

Agregación global y estabilidad

Los valores de Shapley se calcularon en cada iteración del proceso de validación. Para un conjunto representativo de muestras se obtuvo una matriz

$$\hat{\Phi} \in \mathbb{R}^{n_{\text{exp}} \times d}.$$

donde n_{exp} es el número de muestras explicadas (subconjunto X_{exp}) y d es el número de características.

La importancia global de cada característica se resumió mediante el valor medio absoluto de sus contribuciones,

$$I_j = \frac{1}{n_{\text{exp}}} \sum_x |\hat{\phi}_j(x)|,$$

una agregación habitual en prácticas modernas de SHAP. Finalmente, las importancias se normalizaron para obtener contribuciones relativas porcentuales, identificando las variables más relevantes y agrupando el resto en una categoría adicional.

El SHAP personalizado implementado sigue directamente la formulación general de explicaciones por Shapley para modelos de caja negra de [46], aplicada al *pipeline* completo SLMVP-SVM y en el espacio original de características. Como limitación, el apagado progresivo puede generar estados intermedios no observados en los datos reales y el coste computacional es mayor que en métodos SHAP específicos con solución en forma cerrada (p.ej., TreeSHAP o LinearSHAP), al requerir múltiples evaluaciones del modelo por permutaciones; no obstante, no existía una alternativa estándar que ofreciera una explicación igualmente fiel y controlada para este modelo.

Comparación metodológica con KernelSHAP

En principio, la alternativa genérica más inmediata al enfoque utilizado en este trabajo sería KernelSHAP, tratado como un método *model-agnostic*; un método que permite explicar modelos de caja negra aproximando los valores de Shapley mediante muestreo de coaliciones y ajuste de una regresión lineal ponderada local [23]. Sin embargo, la implementación propia explicada anteriormente está más adaptada al *pipeline* SLMVP–SVM, ya que ambas aproximaciones difieren en cómo estiman las contribuciones y en el grado de control que ofrecen sobre el proceso de explicación.

1. **Forma de estimar los valores de Shapley.** KernelSHAP no calcula directamente las contribuciones marginales recorriendo permutaciones completas de las variables, sino que genera coaliciones de características presentes/ausentes y ajusta sobre ellas una regresión lineal ponderada cuyo objetivo es recuperar los valores de Shapley. En cambio, en nuestro método la estimación se realiza de forma explícita mediante permutaciones aleatorias: para cada muestra, se parte de un baseline real, se encienden las características una a una según un orden aleatorio y se acumula el cambio marginal exacto en la salida del modelo en cada paso. Por tanto, la contribución de cada variable se obtiene directamente como promedio de incrementos sucesivos de probabilidad, sin pasar por una regresión intermedia.

2. **Definición del baseline.** En KernelSHAP, las variables ausentes se completan utilizando un conjunto de referencia (*background*), generalmente tratado como un conjunto fijo sobre el que se integran las características faltantes. En la implementación propuesta aquí, el baseline no es fijo, sino **dinámico y específico para cada muestra explicada**: en cada permutación se selecciona aleatoriamente una muestra

real del conjunto de entrenamiento como punto de partida z, excluyendo además la propia muestra que se está explicando. Esto introduce una diferencia metodológica relevante, ya que las atribuciones no se construyen respecto a un único estado medio o genérico, sino respecto a múltiples estados iniciales plausibles observados realmente en los datos de entrenamiento.

3. **Ajuste al *pipeline* completo SLMVP–SVM.** La función explicada en este trabajo no es únicamente una SVM, sino el flujo completo

$$x \longrightarrow \mathit{StandardScaler}(x) \longrightarrow B^{\top}\tilde{x} \longrightarrow \mathit{SVM},$$

es decir, la probabilidad final obtenida tras escalado, proyección supervisada SLMVP y clasificación. Nuestra implementación evalúa siempre este *pipeline* completo, de modo que cada cambio marginal atribuido a una característica recoge simultáneamente su efecto sobre el escalado, sobre la proyección al subespacio y sobre la decisión final de la SVM. Aunque KernelSHAP también podría aplicarse a esta función global tratándola como caja negra, no ofrece por defecto un control tan explícito sobre cómo se generan y recorren los estados perturbados en el espacio original.

4. **Tratamiento de la correlación entre variables.** En radiómica existen con frecuencia dependencias fuertes entre características de intensidad, textura y forma. En nuestro método, el hecho de iniciar cada trayectoria desde una muestra real del conjunto de entrenamiento hace que el estado inicial sea clínicamente plausible y preserve la estructura empírica de correlación de los datos. No obstante, durante el encendido progresivo de variables siguen apareciendo estados intermedios híbridos que pueden no corresponder a observaciones reales.

En consecuencia, la principal ventaja de nuestro enfoque no es cambiar la teoría de Shapley, sino adaptar su aproximación al caso concreto de SLMVP–SVM: las contribuciones se estiman de manera más transparente, con mayor control sobre los baselines y manteniendo el vínculo explícito con las variables radiómicas originales.

3.0.5 Selección de variables para el mejor modelo

Además del *pipeline* completo (107 características), se reentrenó el modelo que proporcionó los mejores resultados, orientado a interpretabilidad y potencial despliegue clínico, restringiendo la entrada a un subconcjunto de las variables originales. Este paso se justifica por dos motivos: (i) el régimen $p \gg n$ (107 variables vs 62 casos) incrementa el riesgo de sobreajuste y la varianza entre particiones, y (ii) un número reducido de descriptores facilita la trazabilidad clínica y la comunicación de la decisión (modelo más *parsimonioso*).

En la práctica clínica, Chowdhury y Turin [8] subrayan que, antes de implementar un modelo predictivo, debe priorizarse la *parsimonia* y el uso de predictores *fácilmente medibles en la práctica clínica*, ya que los modelos con demasiadas variables tienden a ser menos factibles, más costosos de alimentar y con mayor riesgo de falta de datos en condiciones reales [8]. En el contexto de este TFM, donde se parte de 107 descriptores radiómicos extraídos de TC, la reducción a un subconjunto compacto no solo mejora la trazabilidad clínica, sino que también favorece la viabilidad de integración y mantenimiento del modelo en un flujo real, especialmente bajo el régimen $p \gg n$.

El conjunto final de variables se definió como el consenso estable a lo largo de las repeticiones entre los cinco modelos evaluados, evitando que una única partición determine la selección. Con estas variables fijadas, se reentrenó el modelo en $train_t$ y se evaluó en $test_t$, manteniendo el test como conjunto no visto durante la selección.

Capítulo 4

Resultados

4.0.1 Modelos estándar

Todos los modelos se evaluaron siguiendo exactamente el mismo esquema experimental descrito en la sección de Metodología: partición *train–test* 70–30, ajuste interno de hiperparámetros mediante validación cruzada estratificada **optimizando siempre el AUC-ROC** y evaluación final sobre el conjunto de *test*. El procedimiento completo se repitió 15 veces utilizando particiones aleatorias independientes. Las métricas se reportan como media ± desviación estándar entre repeticiones, considerando en cada iteración un conjunto de *test* de 19 casos.

Con respecto a las métricas de evaluación presentadas, el AUC se calculó a partir de las probabilidades estimadas para la clase positiva (glioblastoma), mientras que *accuracy*, *recall*, *precision* y *F1-score* se obtuvieron a partir de las predicciones discretas.

Por su parte, la **matriz de confusión global** presentada en cada modelo se construyó agregando las matrices de confusión obtenidas en cada una de las 15 iteraciones sobre sus respectivos conjuntos de *test*.

Modelo 1: Random Forest

En promedio, el modelo Random Forest alcanzó:

$$\text{Accuracy} = \mathbf{0{,}7158 \pm 0{,}0537}, \quad \text{F1} = \mathbf{0{,}6958 \pm 0{,}0765}, \quad \text{Recall} = \mathbf{0{,}6741 \pm 0{,}1221},$$

$$\text{Precision} = \mathbf{0{,}7368 \pm 0{,}0731}, \quad \text{AUC} = \mathbf{0{,}8074 \pm 0{,}0574}.$$

El resultado de la **matriz de confusión global** se muestra en la Figura 4.1. A partir de esta matriz agregada, correspondiente a un total de $N = 285$ predicciones en test, se derivan los siguientes valores globales: sensibilidad/TPR $= 95/(95 + 46) = 0{,}674$, especificidad/TNR $= 109/(109 + 35) = 0{,}757$ y precisión global $= 95/(95 + 35) = 0{,}731$.

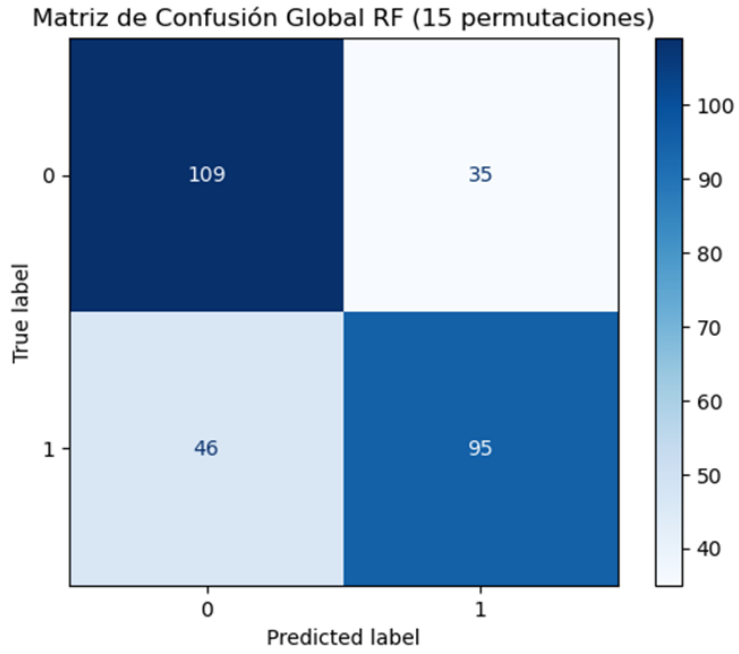


Figura 4.1: Matriz de confusión global de Random Forest (suma de las 15 iteraciones sobre test).

Como ejemplo representativo del material generado por iteración, la Figura 4.2 muestra la curva ROC correspondiente a la Iteración 8 (AUC indicado en la propia figura). Este tipo de visualización se obtuvo para todas las repeticiones, mientras que el valor reportado para el modelo ($0{,}8074 \pm 0{,}0574$) corresponde al promedio $\pm$ desviación estándar del AUC calculado sobre las 15 iteraciones.

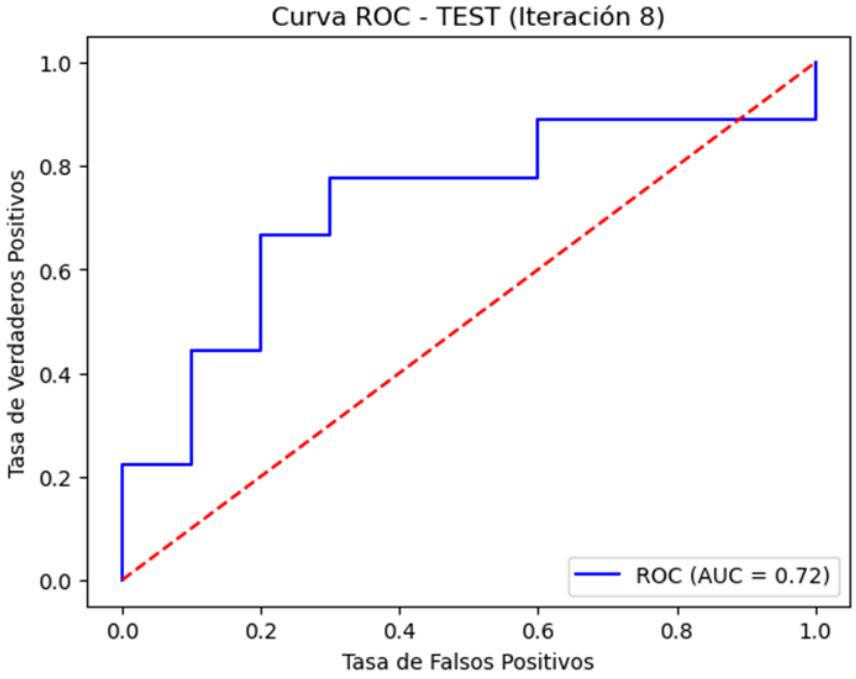


Figura 4.2: Curva ROC en test para Random Forest (ejemplo: Iteración 8).

La Figura 4.3 resume las 10 variables más relevantes según la importancia SHAP media, agrupando el resto de características bajo la categoría *OTRAS_FEATURES*. En este caso, las 10 variables principales concentran conjuntamente el 29,54 % de la importancia total, mientras que el conjunto de variables restantes acumula el 70,46 %. En este caso, para explicar el 50 % del peso de las decisiones del modelo necesitaríamos usar 38 de las 107 *features* originales.

Las tres variables con mayor contribución relativa promedio fueron ***Energy*** (4.39 %), ***ZonePercentage*** (4.09 %) y ***MaximumProbability*** (3.41 %). ***Energy*** es un descriptor de primer orden asociado a la magnitud global de las intensidades de señal, mientras que ***ZonePercentage*** captura la proporción relativa de regiones homogéneas dentro del volumen tumoral, proporcionando información sobre la heterogeneidad espacial de la lesión. Por su parte, ***MaximumProbability***, derivada de matrices de textura, refleja la dominancia de patrones locales específicos de intensidad.

El ranking global incluye descriptores pertenecientes a distintas familias radiómicas, abarcando características de textura (GLCM, GLSZM, GLRLM), estadísticas de primer orden y métricas de forma geométrica (p.ej., *SurfaceVolumeRatio*). La anchura de las distribuciones representadas ("campanas") corresponde a la desviación estándar de la importancia SHAP estimada entre iteraciones, reflejando la variabilidad inducida por el muestreo de los conjuntos de entrenamiento y prueba.

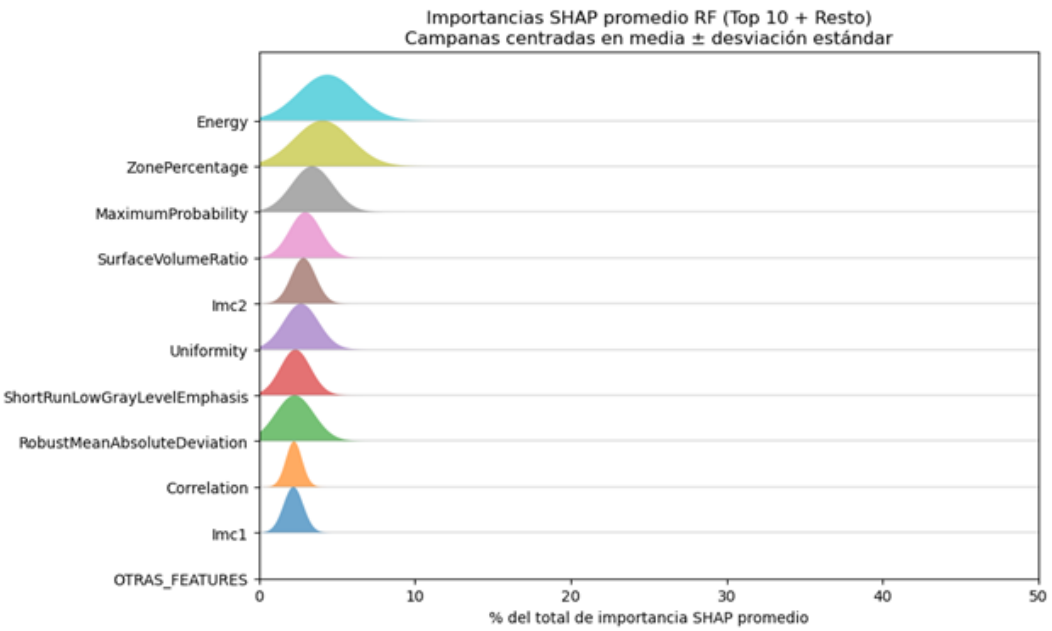
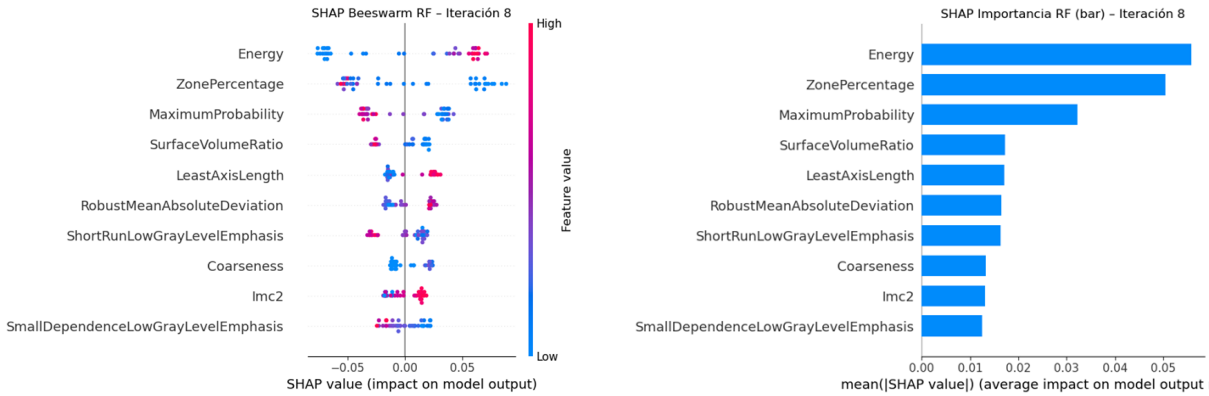


Figura 4.3: Importancias SHAP promediadas en Random Forest (Top 10 + resto), representadas como distribuciones centradas en media ± desviación estándar entre iteraciones.

La Figura 4.4 muestra, para la Iteración 8, (a) el diagrama *beeswarm* y (b) el ranking de importancias basado en mean(|SHAP|). Esta representación permite contrastar la estabilidad del orden de relevancia de las variables a nivel individual frente al patrón promedio observado a lo largo de las repeticiones.



(a) SHAP *beeswarm* (Iteración 8) (b) Importancia según mean(|SHAP|)

Figura 4.4: Explicaciones SHAP para Random Forest (ejemplo: Iteración 8).

Concretamente el gráfico de *beeswarm* permite ver que los valores altos de *Energy* y los valores bajos de *ZonePercentage* y *MaximumProbability* son los que se han asociado con la clase positiva (glioblastoma).

Modelo 2: Explainable Boosting Machine (EBM)

En promedio, el modelo EBM alcanzó:

$$\text{Accuracy} = \mathbf{0{,}7158 \pm 0{,}0632}, \quad \text{F1} = \mathbf{0{,}7025 \pm 0{,}0748}, \quad \text{Recall} = \mathbf{0{,}6896 \pm 0{,}1132},$$

$$\text{Precision} = \mathbf{0{,}7264 \pm 0{,}0606}, \quad \text{AUC} = \mathbf{0{,}7911 \pm 0{,}0587}.$$

El resultado de la **matriz de confusión global** se muestra en la Figura 4.5. A partir de esta matriz agregada, correspondiente a un total de $N = 285$ predicciones en test, se derivan los siguientes valores globales: sensibilidad/TPR $= 97/(97 + 44) = 0{,}688$, especificidad/TNR $= 107/(107 + 37) = 0{,}743$ y precisión global $= 97/(97 + 37) = 0{,}724$.

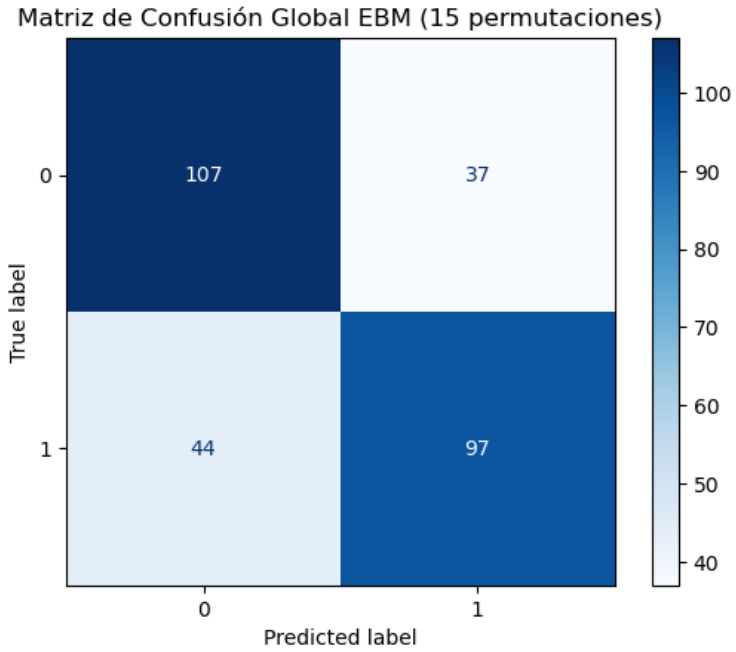


Figura 4.5: Matriz de confusión global de EBM (suma de las 15 iteraciones sobre test).

Como ejemplo representativo del material generado por iteración, la Figura 4.6 muestra la curva ROC correspondiente a la Iteración 14 (AUC indicado en la propia figura). Este tipo de visualización se obtuvo para todas las repeticiones, mientras que el valor reportado para el modelo ($0{,}7911 \pm 0{,}0587$) corresponde al promedio $\pm$ desviación estándar del AUC calculado sobre las 15 iteraciones.

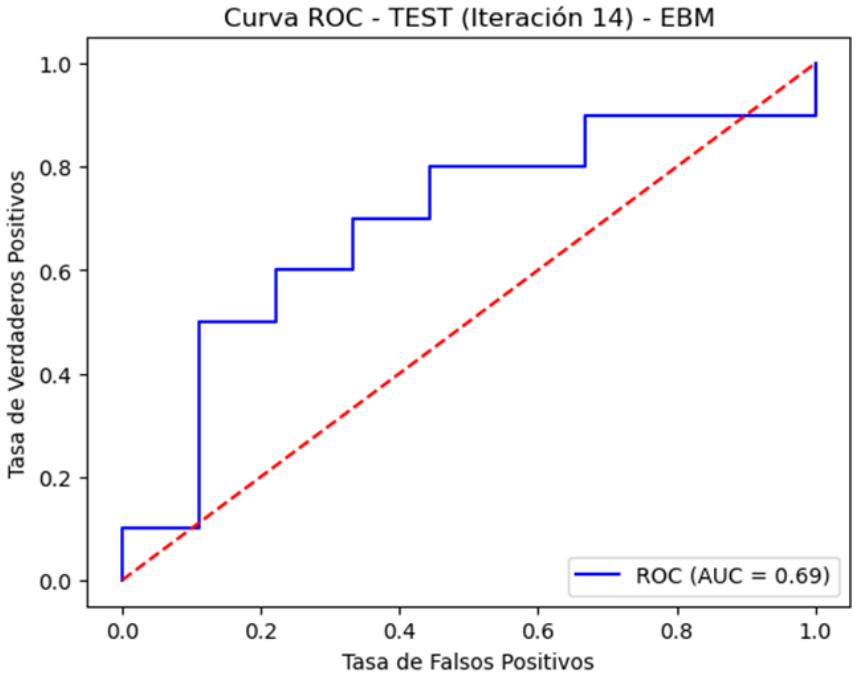


Figura 4.6: Curva ROC en test para EBM (ejemplo: Iteración 14).

La Figura 4.7 resume las 10 variables más relevantes según la importancia SHAP media

(KernelSHAP) y agrupa el resto de características bajo la categoría *OTRAS_FEATURES*. En esta agregación, las 10 variables principales concentran conjuntamente el 16,01 % de la importancia total, mientras que el conjunto de variables restantes acumula el 83,99 %. Debido a este mayor reparto del peso entre las variables para explicar el 50 % del peso de las decisiones del modelo necesitaríamos usar 48 de las 107 *features* originales.

Las tres variables con mayor contribución relativa promedio fueron ***ZonePercentage*** (1.98 %), ***MaximumProbability*** (1.71 %) y ***Energy*** (1.67 %), que en conjunto concentran aproximadamente el 5,36 % de la importancia total. ***ZonePercentage*** (GLSZM) cuantifica la proporción relativa de zonas homogéneas en la lesión, aportando información sobre la heterogeneidad espacial. ***MaximumProbability*** (textura) refleja la probabilidad máxima asociada a patrones locales dominantes de intensidad. Por su parte, ***Energy*** (primer orden) captura la magnitud global de intensidades dentro del volumen segmentado. Entre las variables destacadas también aparecen métricas geométricas como ***MajorAxisLength*** (forma) y descriptores de textura como ***ShortRunLowGrayLevelEmphasis*** (GLRLM), junto con estadísticas de intensidad robustas (p.ej., ***RobustMeanAbsoluteDeviation***) y medidas de dependencia/entropía (p.ej. ***Imc2***).

La anchura de las distribuciones representadas ("campanas") corresponde a la desviación estándar de la importancia SHAP estimada entre iteraciones, reflejando la variabilidad inducida por el muestreo de los conjuntos de entrenamiento y prueba.

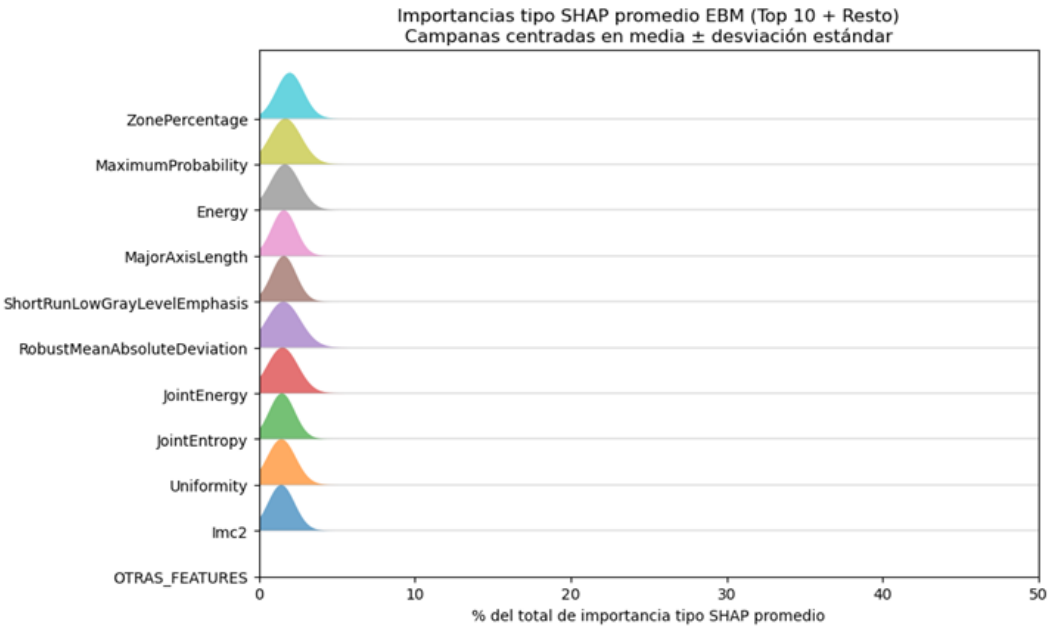
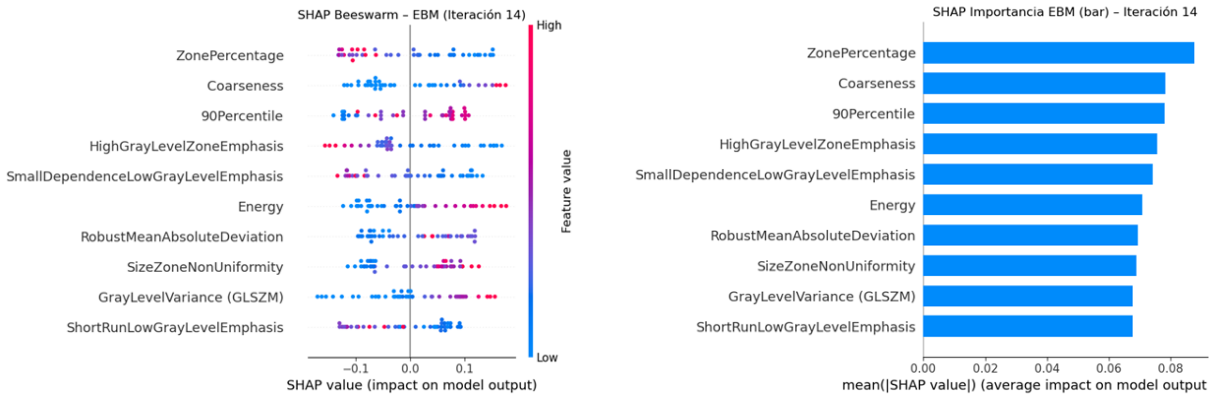


Figura 4.7: Importancias SHAP promediadas en EBM (Top 10 + resto), representadas como distribuciones centradas en media ± desviación estándar entre iteraciones.

La Figura 4.8 muestra, para la Iteración 14, (a) el diagrama *beeswarm* y (b) el ranking de importancias basado en mean(|SHAP|). Esta representación permite analizar el signo/dirección del efecto de cada variable y comparar la estabilidad del orden de relevancia de las *features* en una iteración individual frente al patrón promedio global.



(a) SHAP *beeswarm* (Iteración 14) (b) Importancia según mean(|SHAP|)

Figura 4.8: Explicaciones SHAP para EBM (ejemplo: Iteración 14).

Concretamente el gráfico de *beeswarm* permite ver que los valores altos de *Energy* y los valores bajos de *ZonePercentage* son los que se han asociado con la clase positiva (glioblastoma).

Modelo 3: XGBoost

En promedio, XGBoost alcanzó:

$$\text{Accuracy} = \mathbf{0{,}7263 \pm 0{,}0885}, \quad \text{F1} = \mathbf{0{,}7087 \pm 0{,}0933}, \quad \text{Recall} = \mathbf{0{,}6770 \pm 0{,}1200},$$

$$\text{Precision} = \mathbf{0{,}7628 \pm 0{,}1195}, \quad \text{AUC} = \mathbf{0{,}7889 \pm 0{,}0897}.$$

El resultado de la **matriz de confusión global** se muestra en la Figura 4.9 A partir de esta matriz agregada (sobre $N = 285$ predicciones totales en test sumando iteraciones), se derivan los siguientes valores globales: sensibilidad/TPR $= 95/(95 + 46) = 0{,}674$, especificidad/TNR $= 112/(112 + 32) = 0{,}778$, y precisión global $= 95/(95 + 32) = 0{,}748$

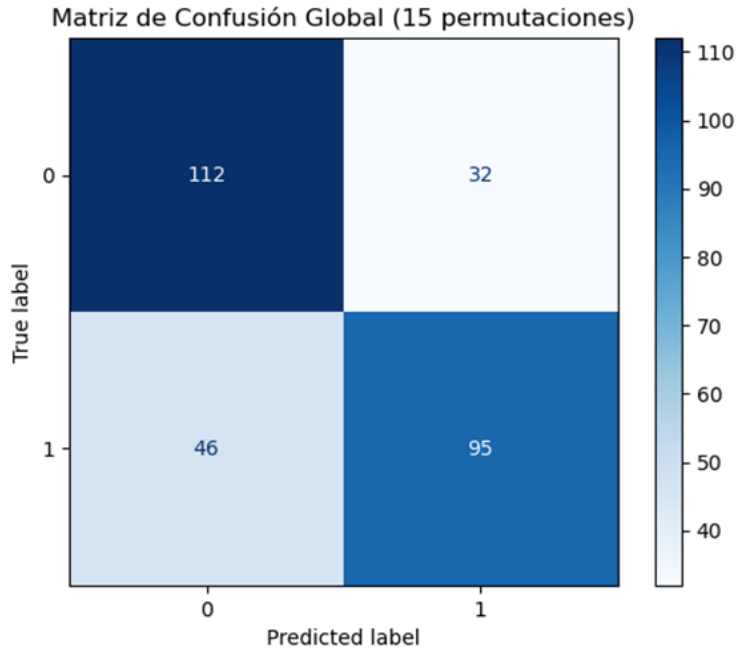


Figura 4.9: Matriz de confusión global de XGBoost (suma de las 15 iteraciones sobre test).

Como ejemplo representativo del material generado por iteración, la Figura 4.10 muestra la curva ROC correspondiente a la Iteración 13 (AUC mostrado en la propia figura). Este tipo de visualización se obtuvo para todas las repeticiones, mientras que el valor reportado para el modelo ($0{,}7889 \pm 0{,}0897$) corresponde al promedio $\pm$ desviación estándar del AUC sobre las 15 iteraciones.

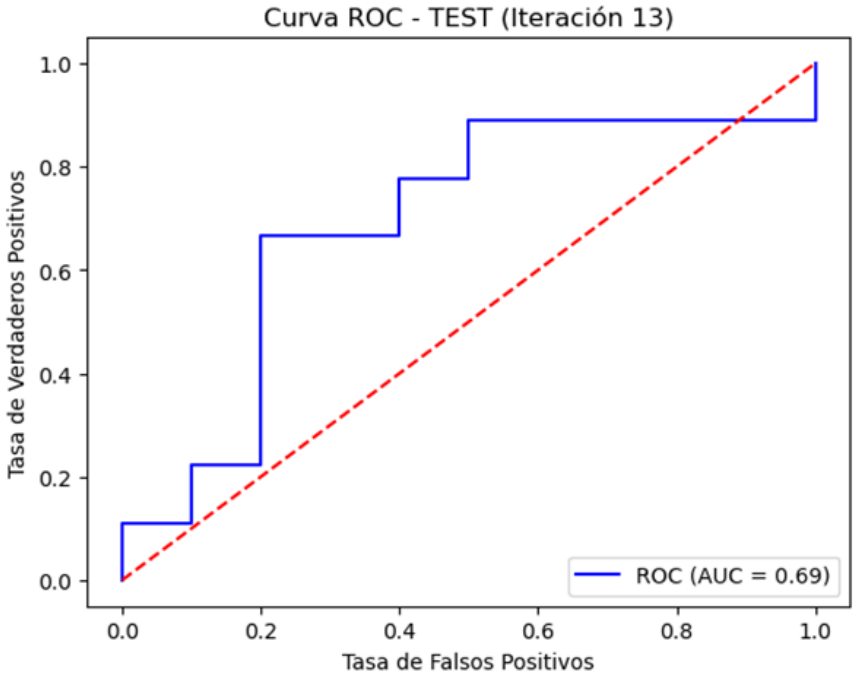


Figura 4.10: Curva ROC en test para XGBoost (ejemplo: Iteración 13).

La Figura 4.11 resume las 10 variables más discriminativas según importancia media

y agrupa el resto como *OTRAS_FEATURES*. En esta agregación, estas 10 variables concentran el 59,91 % de la importancia total, mientras que el resto acumulan el 40,09 % restante.

Las tres variables con mayor contribución relativa promedio fueron ***ZonePercentage*** (13.64 %), ***Energy*** (12.07 %) y ***MaximumProbability*** (9.44 %), que en conjunto concentran aproximadamente el 35,15 % de la importancia total. Para explicar el 50 % del peso de las decisiones del modelo necesitaríamos usar tan sólo 7 de las 107 *features* originales. ***ZonePercentage*** cuantifica la proporción relativa de regiones homogéneas dentro del volumen tumoral, capturando información asociada a la heterogeneidad espacial de la lesión. Por su parte, ***Energy*** es un descriptor de primer orden que mide la magnitud global de las intensidades de señal, estando relacionado con la uniformidad y el nivel energético del tejido analizado. Finalmente, ***MaximumProbability***, derivada de matrices de textura, refleja la probabilidad máxima de ocurrencia de un patrón de intensidades concreto, aportando información sobre la dominancia de determinadas estructuras locales.

En términos generales, el ranking incluye descriptores pertenecientes a distintas familias radiómicas, abarcando características de textura (p.ej., GLCM, GLSZM, GLRLM), estadísticas de primer orden y métricas de forma geométrica (p.ej., *MajorAxisLength*, *SurfaceVolumeRatio*). La anchura de las distribuciones representadas ("campanas") corresponde a la desviación estándar de la importancia estimada entre iteraciones, reflejando la variabilidad asociada al muestreo de los conjuntos de entrenamiento y prueba.

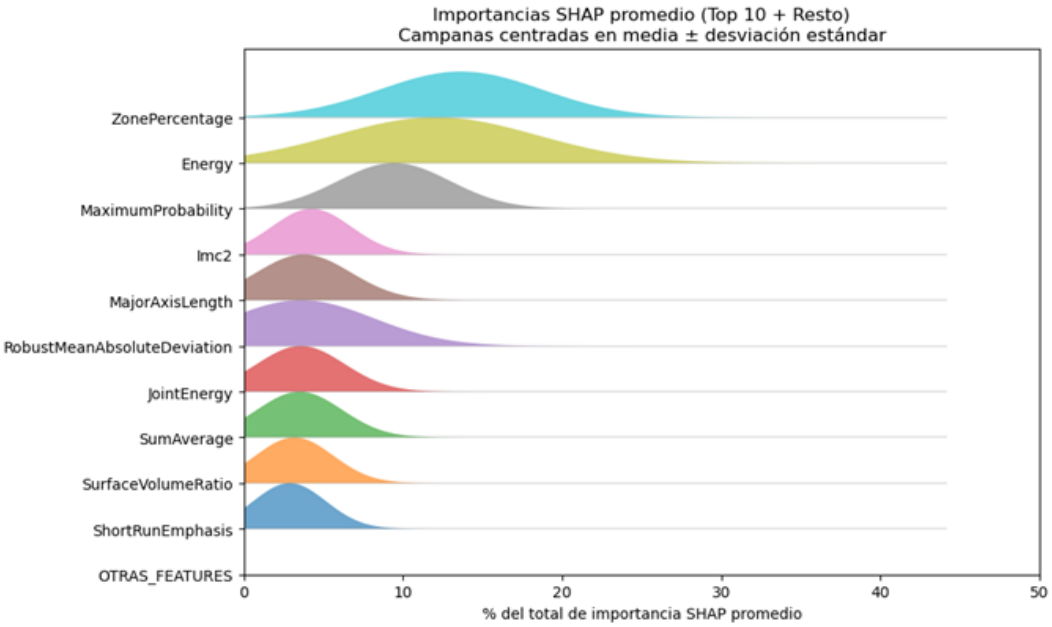
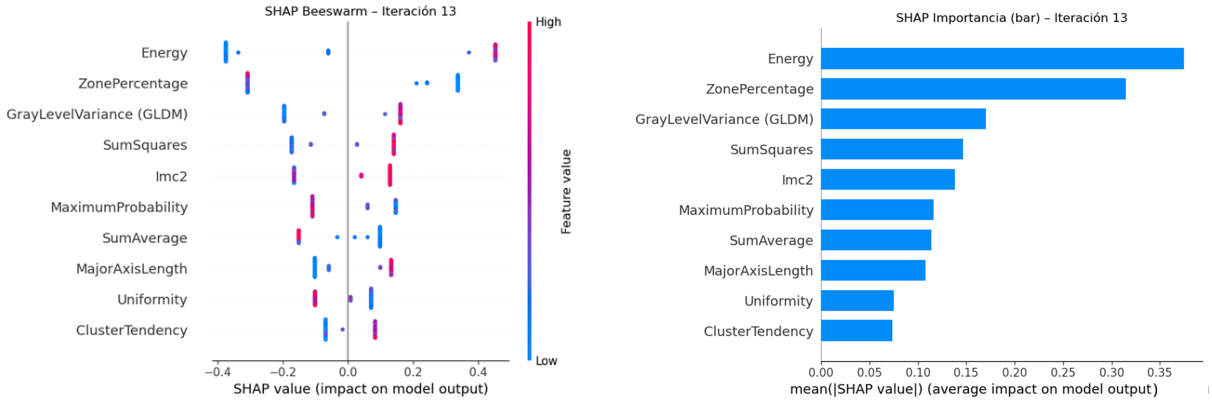


Figura 4.11: Importancias SHAP promediadas en XGBoost (Top 10 + resto), representadas como distribuciones centradas en media ± desviación estándar entre iteraciones.

La Figura 4.12 muestra, para la Iteración 13, (a) el diagrama *beeswarm* y (b) el ranking de importancias según mean(|SHAP|). Esta representación conjunta permite comparar visualmente la estabilidad del orden de relevancia de las variables entre iteraciones individuales y el promedio global.



(a) SHAP *beeswarm* (Iteración 13) (b) Importancia según mean(|SHAP|)

Figura 4.12: Explicaciones SHAP para XGBoost (ejemplo: Iteración 13).

Concretamente el gráfico de *beeswarm* permite ver que los valores altos de *Energy* y los valores bajos de *ZonePercentage* y *MaximumProbability* son los que se han asociado con la clase positiva (glioblastoma).

Modelo 4: Support Vector Machine (SVM)

En promedio, el modelo SVM alcanzó:

$$\text{Accuracy} = \mathbf{0{,}7263 \pm 0{,}0774}, \quad \text{F1} = \mathbf{0{,}6789 \pm 0{,}1133}, \quad \text{Recall} = \mathbf{0{,}6111 \pm 0{,}1556},$$

$$\text{Precision} = \mathbf{0{,}8017 \pm 0{,}1069}, \quad \text{AUC} = \mathbf{0{,}8181 \pm 0{,}0609}.$$

El resultado de la **matriz de confusión global** se muestra en la Figura 4.13. A partir de esta matriz agregada, correspondiente a un total de $N = 285$ predicciones en test, se derivan los siguientes valores globales: sensibilidad/TPR $= 86/(86 + 55) = 0{,}611$, especificidad/TNR $= 121/(121 + 23) = 0{,}840$ y precisión global $= 86/(86 + 23) = 0{,}789$.

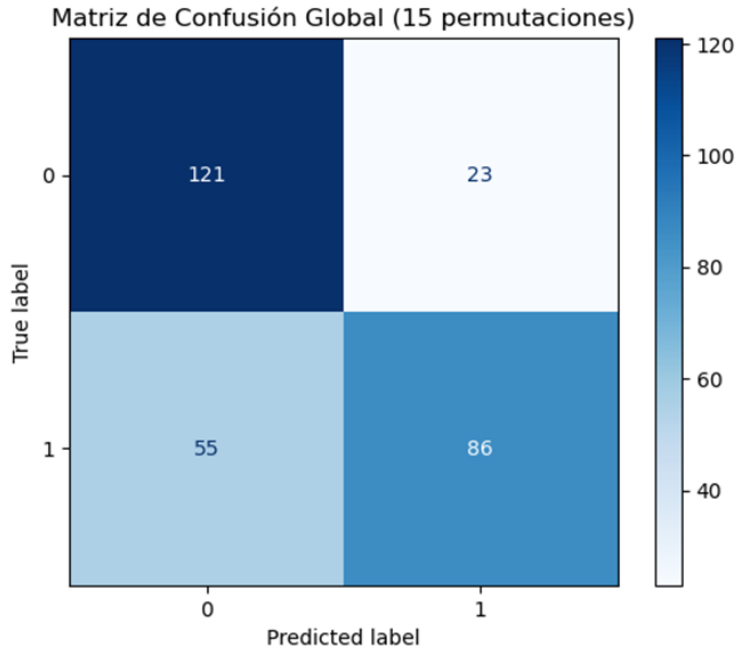


Figura 4.13: Matriz de confusión global de SVM (suma de las 15 iteraciones sobre test).

Como ejemplo representativo del material generado por iteración, la Figura 4.14 muestra la curva ROC correspondiente a la Iteración 4 (AUC indicado en la propia figura). Este tipo de visualización se obtuvo para todas las repeticiones, mientras que el valor reportado para el modelo ($0{,}8181 \pm 0{,}0609$) corresponde al promedio $\pm$ desviación estándar del AUC calculado sobre las 15 iteraciones.

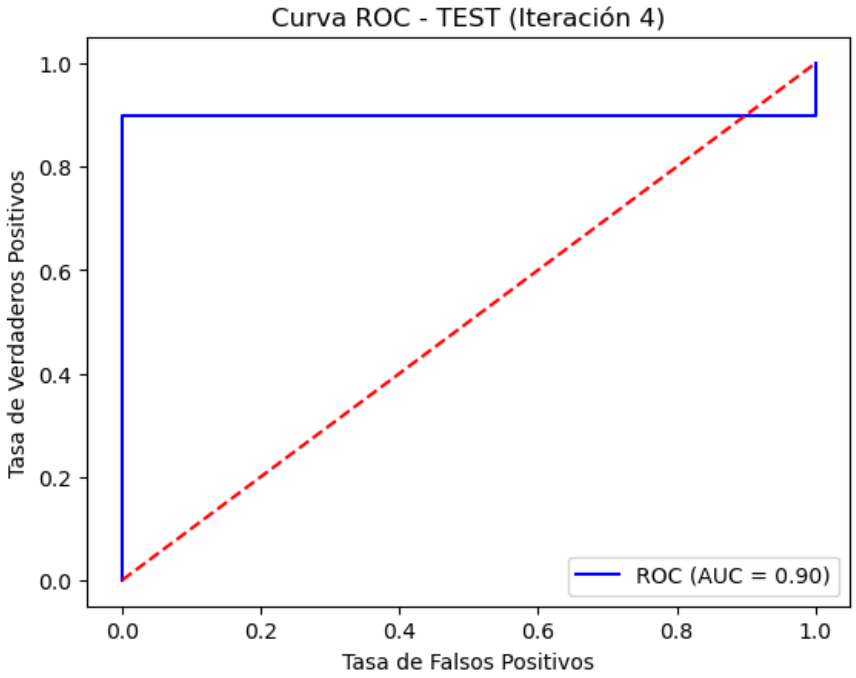


Figura 4.14: Curva ROC en test para SVM (ejemplo: Iteración 4).

La Figura 4.15 resume las 10 variables más relevantes según la importancia SHAP media

y agrupa el resto de características bajo la categoría *OTRAS_FEATURES*. En esta agregación, estas 10 variables concentran el 22,52 % de la importancia total, mientras que el conjunto de variables restantes acumula el 77,48 %, lo que refleja una distribución de la relevancia más dispersa que en los modelos basados en *boosting*.

Las tres variables con mayor contribución relativa promedio fueron ***Energy*** (2.86 %), ***Imc1*** (2.51 %) y ***GrayLevelNonUniformityNormalized (GLSZM)*** (2.50 %), que en conjunto concentran aproximadamente el 7,87 % de la importancia total. Para explicar el 50 % del peso de las decisiones del modelo necesitaríamos usar 45 de las 107 *features* originales. ***Energy*** es un descriptor de primer orden asociado a la magnitud global de las intensidades dentro del volumen tumoral. ***Imc1***, derivada de matrices GLCM, captura dependencias informacionales entre niveles de gris, relacionadas con la complejidad textural de la lesión. Por su parte, ***GrayLevelNonUniformityNormalized (GLSZM)*** cuantifica la variabilidad de los niveles de gris en zonas homogéneas, aportando información sobre la irregularidad interna del tejido.

En términos generales, el ranking incluye descriptores pertenecientes a distintas familias radiómicas, abarcando estadísticas de primer orden, características de textura basadas en GLCM y GLSZM, así como métricas geométricas de forma (p.ej., *SurfaceVolumeRatio*). La anchura de las distribuciones representadas ("campanas") corresponde a la desviación estándar de la importancia SHAP estimada entre iteraciones, reflejando la variabilidad asociada al muestreo de los conjuntos de entrenamiento y prueba.

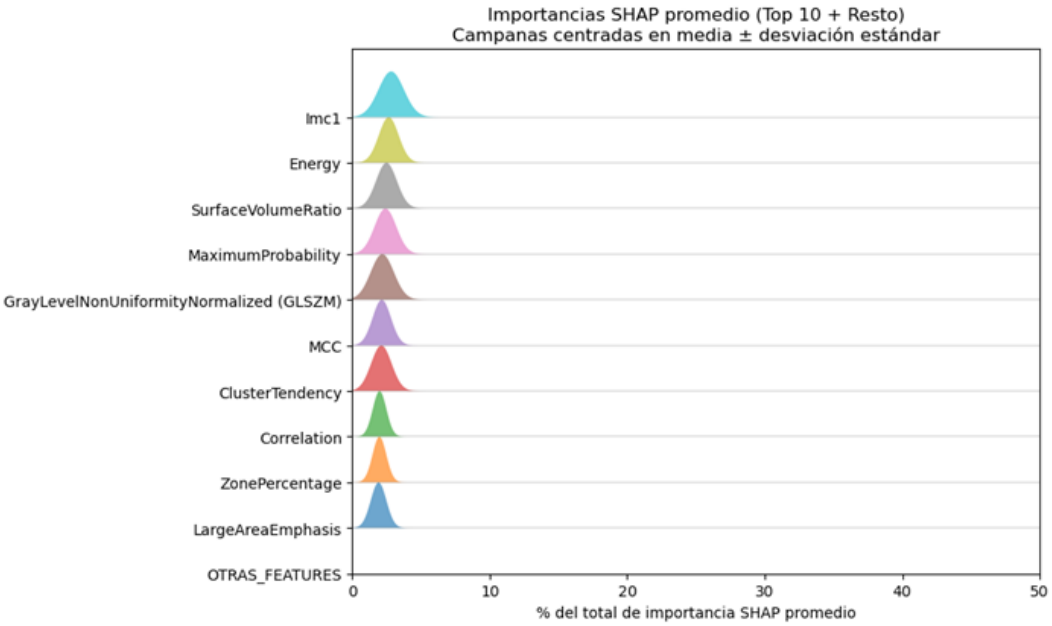
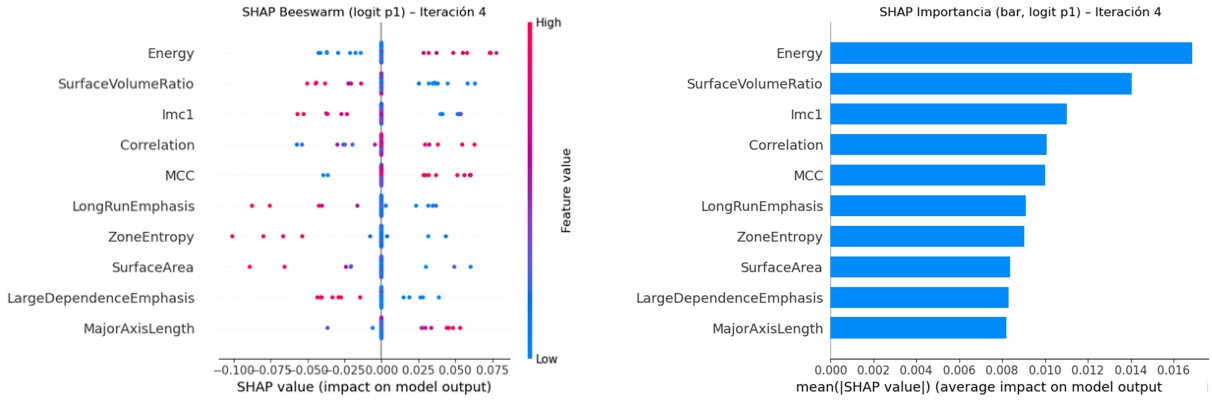


Figura 4.15: Importancias SHAP promediadas en SVM (Top 10 + resto), representadas como distribuciones centradas en media ± desviación estándar entre iteraciones.

La Figura 4.16 muestra, para la Iteración 4, (a) el diagrama *beeswarm* y (b) el ranking de importancias basado en mean(|SHAP|). Esta representación permite analizar el comportamiento local del modelo y contrastar la coherencia del orden de relevancia de las variables en una iteración individual con el patrón promedio observado globalmente.



(a) SHAP *beeswarm* (Iteración 4) (b) Importancia según mean(|SHAP|)

Figura 4.16: Explicaciones SHAP para SVM (ejemplo: Iteración 4).

Concretamente el gráfico de *beeswarm* permite ver que los valores altos de *Energy* y los valores bajos de *SurfaceVolumeRatio* y de *Imc1* son los que se han asociado con la clase positiva (glioblastoma).

4.0.2 SLMVP-SVM

Selección del *rank* (número de proyecciones)

Un parámetro crítico del modelo SLMVP–SVM es la dimensión del subespacio proyectado, o *rank* r, que determina cuántas componentes supervisadas se conservan tras la proyección. En este problema (107 *features* y $n = 62$ casos), valores de r demasiado bajos pueden comprimir en exceso la señal discriminativa, mientras que valores altos tienden a reintroducir ruido y redundancia, aumentando el riesgo de sobreajuste y la variabilidad entre particiones.

Por este motivo, antes de fijar r en el *pipeline* final, se estudió cómo evolucionaba el rendimiento del modelo en función del número de proyecciones. En concreto, se evaluó un rango amplio de valores r, desde proyecciones muy compactas hasta el valor máximo admisible, que viene determinado por el rango efectivo de la matriz de proyección aprendida en SLMVP. En este caso, dicho rango queda acotado por $\text{mín}(n_{\text{train}} - 1, d)$, donde n_{train} es el número de muestras disponibles en el conjunto de entrenamiento y d la dimensionalidad original del espacio de características.

Dado que en cada iteración se emplearon $n_{\text{train}} = 43$ casos de entrenamiento y $d = 107$ características radiómicas, el número máximo de proyecciones linealmente independientes quedó limitado a $r_{\text{máx}} = 42$. En consecuencia, el análisis de sensibilidad del rendimiento se realizó barriendo valores de $r \in [1, 42]$, evaluando para cada uno de ellos el comportamiento del *pipeline* completo SLMVP–SVM bajo el mismo esquema de validación descrito en la sección de Metodología.

Para cada r, se entrenó el flujo completo SLMVP–SVM siguiendo el mismo esquema de evaluación descrito en Metodología (partición 70–30 con ajuste interno mediante validación cruzada estratificada), registrando el *accuracy* obtenido en *test*.

En la Figura 4.17 se observa que para valores muy bajos de r, el rendimiento es inestable y claramente inferior, indicando una proyección excesivamente restrictiva que no preserva suficiente información discriminativa. A partir de valores intermedios bajos, el *accuracy*

aumenta de forma abrupta y entra en una meseta amplia, con rendimientos elevados y relativamente estables en un intervalo aproximado $r \in [5, 30]$.

Más allá de este rango, el rendimiento muestra una tendencia progresiva a la degradación conforme aumenta el número de proyecciones. Este comportamiento es hasta cierto punto esperable, ya que sugiere que incorporar dimensiones adicionales introduce ruido y redundancia en el subespacio proyectado, lo que resulta especialmente perjudicial en un escenario de baja dimensionalidad.

En base a este análisis, se seleccionó $r = \mathbf{10}$ ya que fue este valor para el que se obtuvieron mejores resultados. Todo el análisis de explicabilidad y métricas de evaluación reportadas a partir de este punto para el clasificador SLMVP–SVM se obtienen fijando dicho valor de *rank*.

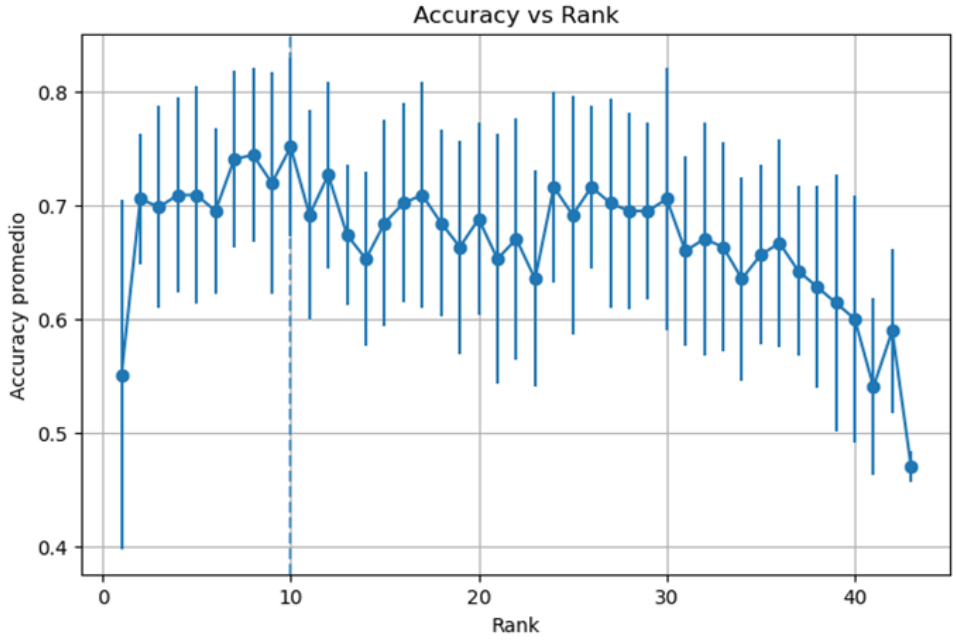


Figura 4.17: Evolución del *accuracy* en *test* en función del *rank* r (número de proyecciones) en SLMVP–SVM. La selección final $r = 10$ se realizó por saturación del rendimiento y criterio de parsimonia.

Resultados del modelo

En promedio, el modelo SLMVP–SVM alcanzó:

$$\text{Accuracy} = \mathbf{0{,}7509 \pm 0{,}0780}, \quad \text{F1} = \mathbf{0{,}7430 \pm 0{,}0899}, \quad \text{Recall} = \mathbf{0{,}7459 \pm 0{,}1368},$$

$$\text{Precision} = \mathbf{0{,}7574 \pm 0{,}0862}, \quad \text{AUC} = \mathbf{0{,}8085 \pm 0{,}0749}.$$

El AUC se calculó a partir de las probabilidades estimadas para la clase positiva (glioblastoma), mientras que *accuracy*, *recall*, *precision* y *F1-score* se obtuvieron a partir de las predicciones discretas del clasificador.

El resultado de la **matriz de confusión global** se muestra en la Figura 4.18. A partir de esta matriz agregada, correspondiente a un total de $N = 285$ predicciones en test, se derivan los siguientes valores globales: sensibilidad/TPR $= 105/(105 + 36) = 0{,}745$, especificidad/TNR $= 109/(109 + 35) = 0{,}757$ y precisión global $= 105/(105 + 35) = 0{,}750$.

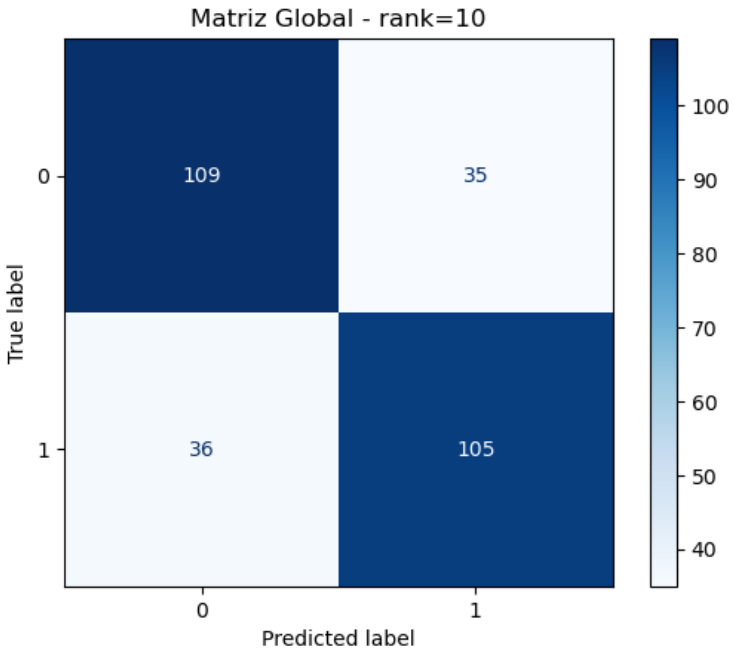


Figura 4.18: Matriz de confusión global de SLMVP–SVM (suma de las 15 iteraciones sobre test).

Como ejemplo representativo del material generado por iteración, la Figura 4.19 muestra la curva ROC correspondiente a la Iteración 4 (AUC indicado en la propia figura). Este tipo de visualización se obtuvo para todas las repeticiones, mientras que el valor reportado para el modelo ($0{,}8085 \pm 0{,}0749$) corresponde al promedio $\pm$ desviación estándar del AUC sobre las 15 iteraciones.

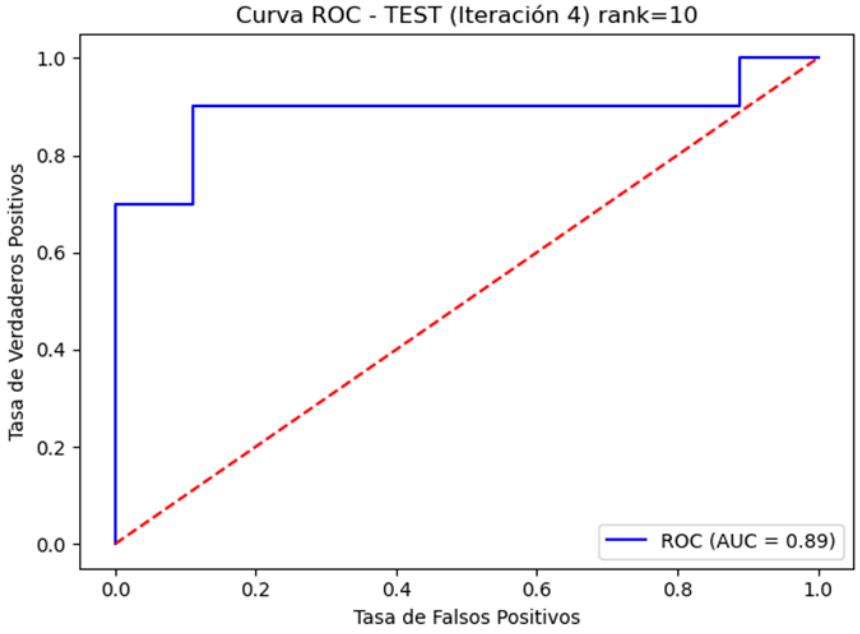


Figura 4.19: Curva ROC en test para SLMVP–SVM (ejemplo: Iteración 4).

La Figura 4.20 resume las 10 variables más relevantes según la importancia SHAP promedio

(implementación *custom* descrita en la sección de Explicabilidad), agrupando el resto de características bajo la categoría *OTRAS_FEATURES*. En esta agregación, las 10 variables principales concentran conjuntamente el 18,60 % de la importancia total, mientras que el conjunto de variables restantes acumula el 81,40 %. En este caso, para explicar el 50 % del peso de las decisiones del modelo necesitaríamos usar 48 de las 107 *features* originales.

Las tres variables con mayor contribución relativa promedio fueron ***Energy*** (2.43 %), ***LeastAxisLength*** (2.01 %) y ***LargeAreaLowGrayLevelEmphasis*** (1.96 %). ***Energy*** es un descriptor de primer orden asociado a la magnitud global de las intensidades dentro del volumen segmentado. ***LeastAxisLength*** es una característica de forma que refleja la extensión mínima de la lesión, aportando información geométrica complementaria a la esfericidad o elongación. Por su parte, ***LargeAreaLowGrayLevelEmphasis*** (familia de textura, GLSZM) cuantifica la presencia relativa de zonas extensas con niveles de gris bajos, capturando patrones de homogeneidad/heterogeneidad asociados a la estructura interna de la lesión.

En términos generales, el ranking incluye descriptores de primer orden, forma y textura (GLCM/GLSZM/GLRLM). La anchura de las distribuciones representadas ("campanas") corresponde a la desviación estándar de la importancia SHAP estimada entre iteraciones, reflejando la variabilidad inducida por el muestreo de los conjuntos de entrenamiento y prueba.

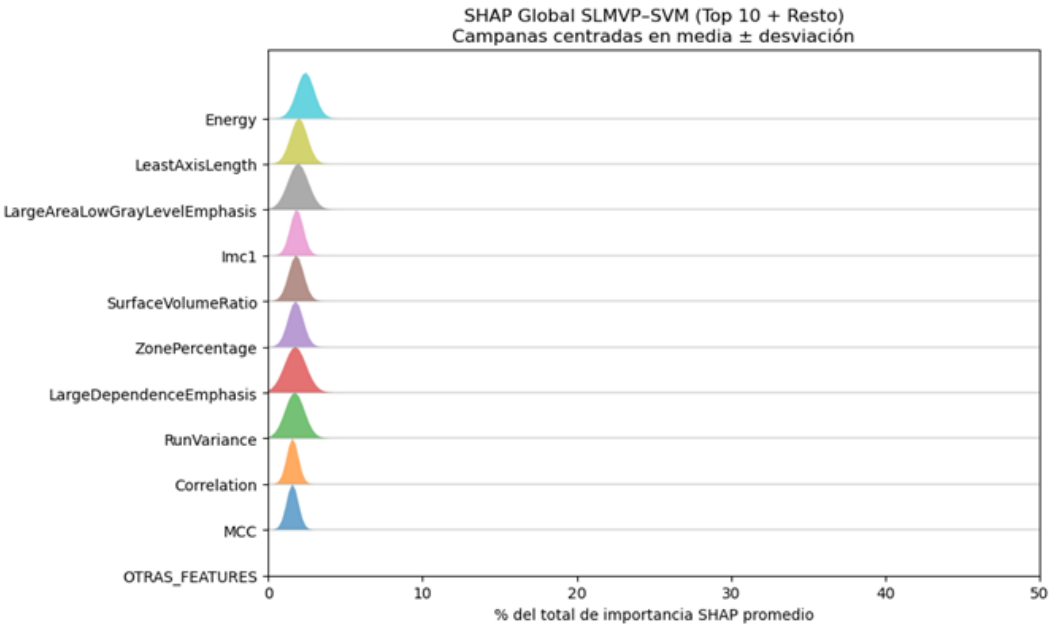
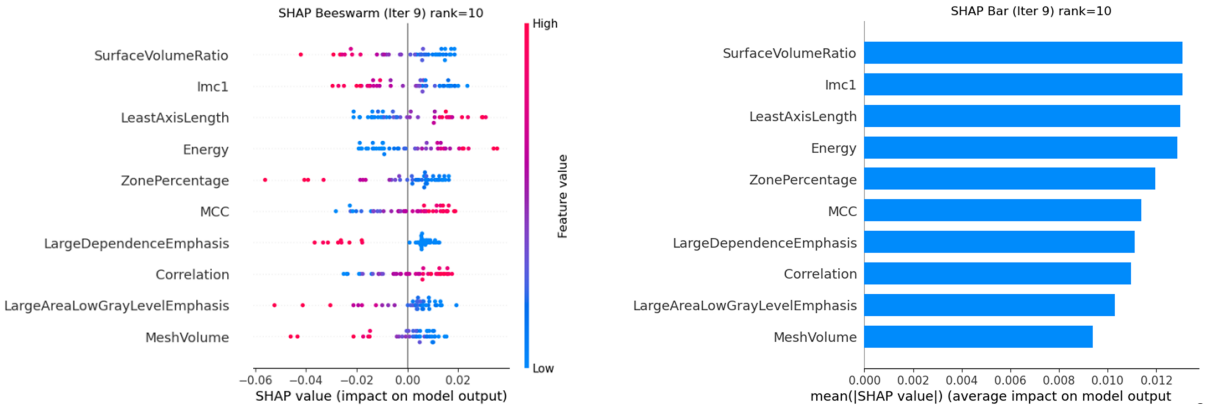


Figura 4.20: Importancias SHAP promediadas en SLMVP–SVM (Top 10 + resto), representadas como distribuciones centradas en media ± desviación estándar entre iteraciones.

La Figura 4.21 muestra, para la Iteración 9, (a) el diagrama *beeswarm* y (b) el ranking de importancias basado en mean(|SHAP|). Esta representación permite analizar el comportamiento local del *pipeline* completo (estandarización + SLMVP + SVM) y contrastar la coherencia del orden de relevancia de las variables en una iteración individual con el patrón promedio global.



(a) SHAP *beeswarm* (Iteración 9) (b) Importancia según mean(|SHAP|)

Figura 4.21: Explicaciones SHAP para SLMVP–SVM (ejemplo: Iteración 9).

Concretamente el gráfico de *beeswarm* permite ver que los valores altos de *Energy* y de *LeastAxisLength* y bajos de *LargeAreaLowGrayLevelEmphasis* son los que se han asociado con la clase positiva (glioblastoma).

4.0.3 Comparación entre modelos

Con el objetivo de facilitar una comparación directa entre los resultados obtenidos, el Cuadro 4.1 resume las métricas promedio (± desviación estándar) obtenidas en las 15 repeticiones para los cinco clasificadores evaluados. En conjunto, los resultados muestran diferencias moderadas en rendimiento medio, pero diferencias relevantes tanto en el *perfil de error* (sensibilidad vs especificidad) como en la *estabilidad* entre particiones, un aspecto crítico dado el tamaño limitado del conjunto de test por iteración ($n_{\text{test}} = 19$).

Modelo	Accuracy	F1	Recall	Precision	AUC
Random Forest	$0{,}715 \pm 0{,}053$	$0{,}6958 \pm 0{,}076$	$0{,}6741 \pm 0{,}122$	$0{,}736 \pm 0{,}073$	$0{,}807 \pm 0{,}057$
EBM	$0{,}715 \pm 0{,}063$	$0{,}702 \pm 0{,}074$	$0{,}689 \pm 0{,}113$	$0{,}726 \pm 0{,}060$	$0{,}791 \pm 0{,}058$
XGBoost	$0{,}726 \pm 0{,}088$	$0{,}708 \pm 0{,}093$	$0{,}677 \pm 0{,}120$	$0{,}762 \pm 0{,}119$	$0{,}788 \pm 0{,}089$
SVM	$0{,}726 \pm 0{,}077$	$0{,}678 \pm 0{,}113$	$0{,}611 \pm 0{,}155$	$\mathbf{0{,}801 \pm 0{,}106}$	$\mathbf{0{,}818 \pm 0{,}060}$
SLMVP–SVM	$\mathbf{0{,}750 \pm 0{,}078}$	$\mathbf{0{,}743 \pm 0{,}089}$	$\mathbf{0{,}745 \pm 0{,}136}$	$0{,}757 \pm 0{,}086$	$0{,}808 \pm 0{,}074$

Tabla 4.1: Resumen comparativo de rendimiento (media ± desviación estándar) en test a lo largo de 15 repeticiones.

En términos de *accuracy* y *F1*, el modelo SLMVP–SVM presenta la mejor media global, junto con una sensibilidad (recall) claramente superior. Esta diferencia es clínicamente relevante si se prioriza minimizar falsos negativos de glioblastoma (clase positiva). En contraste, la SVM estándar muestra el mayor AUC medio y la mayor precisión, coherente con una frontera más conservadora (menos falsos positivos) pero a costa de una menor sensibilidad (más falsos negativos). Los modelos basados en árboles (RF y XGBoost) y el modelo aditivo interpretable (EBM) quedan en una zona intermedia: rendimientos medios

similares entre sí y perfiles de error más equilibrados que la SVM, aunque sin alcanzar el recall de SLMVP–SVM.

La estabilidad puede analizarse a dos niveles: (i) dispersión de métricas (desviación estándar) y (ii) dispersión de importancias globales (las "campanas" en las figuras SHAP promediadas). En el primer nivel, Random Forest muestra la menor desviación estándar en *accuracy*, mientras que XGBoost exhibe la mayor variabilidad tanto en *accuracy* como en AUC, consistente con una mayor sensibilidad del *boosting* a cambios en el subconjunto de entrenamiento/test cuando el tamaño muestral es bajo. En el segundo nivel, XGBoost tiende a concentrar gran parte del peso en pocas variables (concentración del $\approx 60\,\%$ del peso de las explicaciones en las 10 características más relevantes) aunque con mucha desviación, lo que muestra grandes diferencias entre iteraciones; por el contrario, EBM, SVM, y SLMVP-SVM reparten la importancia entre muchas *features* ($\approx 16\,\%{-}23\,\%$), generando "campanas" menos dispersas y cercanas entre sí, generando así un ranking más sensible a correlaciones y al método de aproximación (KernelSHAP).

Los *beeswarm* difieren de forma sistemática entre familias de modelos. En XGBoost, es habitual observar puntos agrupados en bandas "columnares" (valores SHAP cuantizados), porque las contribuciones de TreeSHAP se obtienen como sumas de saltos discretos ligados a *splits* y hojas; con hiperparámetros que restringen el crecimiento del árbol (p.,ej., `max_depth`, `min_child_weight`, `gamma` y la regularización), se reducen las trayectorias efectivas y muchas muestras acaban compartiendo rutas similares, acumulándose en valores SHAP repetidos. En Random Forest este patrón suele ser menos marcado al promediarse muchos árboles construidos con *bootstrap* y aleatoriedad de variables. En cambio, en SVM (KernelSHAP) se observa con frecuencia un patrón "colapsado" alrededor de cero en varias variables, que puede explicarse por varios factores complementarios:

(i) **Frontera de decisión suave del SVM:** especialmente cuando se emplean kernels no lineales y una posterior calibración probabilística, la función de decisión es continua y presenta transiciones graduales. Como consecuencia, al perturbar una característica (encenderla o apagarla en el sentido de KernelSHAP), el cambio marginal en la salida del modelo suele ser reducido, lo que se traduce en valores SHAP cercanos a cero para muchas variables.

(ii) **Reparto de contribución por colinealidad:** cuando existen variables altamente correlacionadas, KernelSHAP tiende a distribuir la contribución total entre todas ellas. Este reparto diluye el impacto individual de cada característica, reduciendo el valor absoluto de los SHAP asociados a cada una y favoreciendo la aparición de concentraciones alrededor del valor base.

A pesar de estas diferencias entre los modelos, existe un núcleo de *features* recurrentes (aparecen en ≥ 3 modelos entre las 10 más relevantes): *ZonePercentage*, *Energy*, *Maximum-Probability*, *Imc2*, *SurfaceVolumeRatio*, *RobustMeanAbsoluteDeviation*, *Imc1 Correlation*. La Figura 4.22 representa su importancia media (en porcentaje) y su dispersión (desviación estándar) agregada entre modelos, actuando como una síntesis de consistencia inter-modelo.

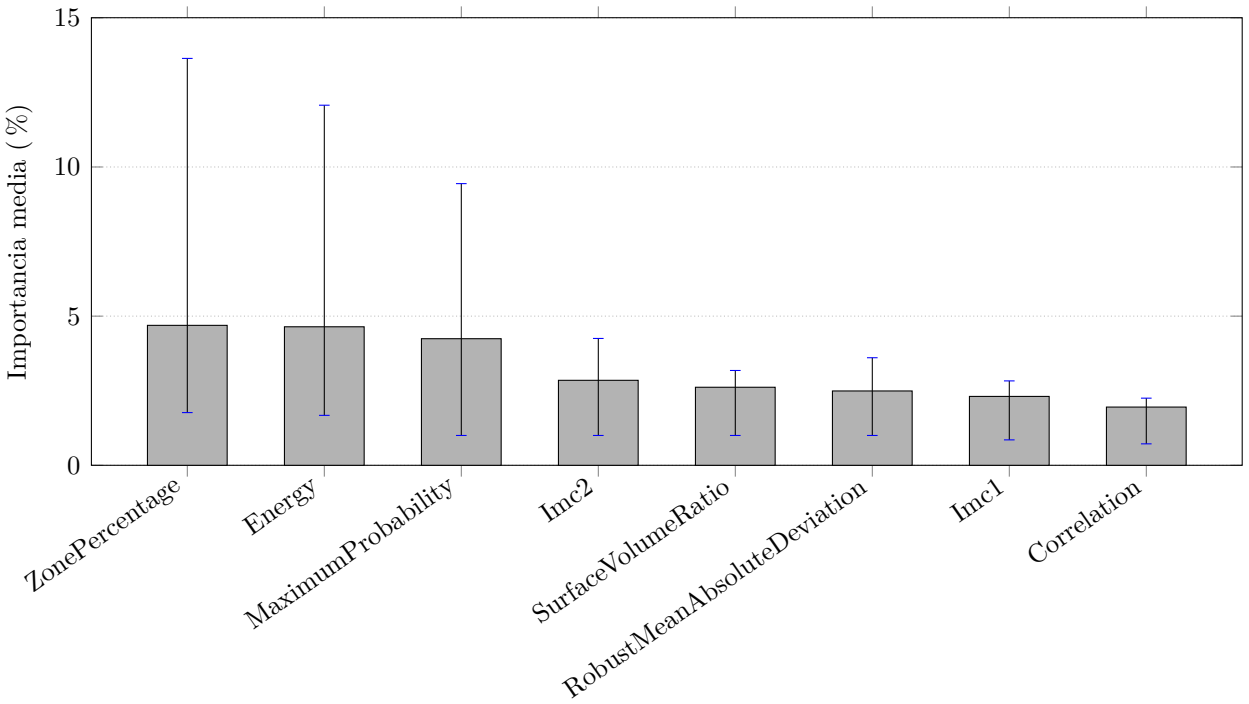


Figura 4.22: Síntesis de *features* que aparecen entre las 10 más relevantes en ≥ 3 modelos. Se muestra importancia media (%) y rango (mín–máx)

Además de este análisis basado en la media de importancias, se propuso un enfoque alternativo para comparar los resultados que no depende únicamente del valor medio de relevancia de cada métrica. En su lugar, se evalúa la concordancia entre los rankings de variables más influyentes reportados por cada modelo frente a un ranking consenso.

Dado que en problemas de explicabilidad no existe un *ground truth* natural sobre cuáles son las variables verdaderamente más importantes para el modelo, resulta necesario definir una referencia operativa que permita comparar de forma sistemática los rankings obtenidos por distintos clasificadores. En este sentido, [21] señalan que, ante la ausencia de explicaciones de referencia, una estrategia razonable consiste en construir un *pseudo ground truth* a partir del consenso entre múltiples modelos bien entrenados. La idea subyacente es que, si varias arquitecturas diferentes destacan de forma recurrente un mismo subconjunto de variables, dichas variables constituyen una aproximación más robusta a la señal explicativa compartida que la interpretación derivada de un único modelo o de una única partición de los datos.

Siguiendo esta lógica, en este trabajo se construyó un ranking consenso agregando la información procedente de los cinco modelos evaluados. Concretamente, se priorizó la frecuencia de aparición de cada variable en los rankings de importancia, incorporando además como criterio secundario su porcentaje medio de contribución relativa. De este modo, el consenso no debe interpretarse como una verdad absoluta sobre la relevancia de las características, sino como una referencia agregada y estable —análogo al *pseudo ground truth* planteado por [21]— que permite cuantificar hasta qué punto los rankings individuales de cada modelo convergen hacia una estructura explicativa común.

Construcción del consenso.

Sea $\mathcal{M}$ el conjunto de modelos evaluados. Cada modelo $m \in \mathcal{M}$ aporta un ranking r_m con

sus 10 variables más importantes. Se define el universo de variables como la unión

$$\mathcal{F} = \bigcup_{m \in \mathcal{M}} r_m.$$

A cada variable $f \in \mathcal{F}$ se le asigna una puntuación de consenso basada en: (i) su frecuencia de aparición entre las 10 características más relevantes en cada modelo, y (ii) un desempate suave mediante el porcentaje relativo de contribución (% del total). Formalmente,

$$s(f) = \text{freq}(f) + \lambda \cdot \text{norm}(\overline{p}(f)),$$

donde $\text{freq}(f)$ es el número de modelos cuyo ranking incluye f, $\overline{p}(f)$ es el promedio del % del total (si existe) y $\text{norm}(\cdot)$ es una normalización min–max a $[0, 1]$ sobre $\{\overline{p}(f)\}_{f \in \mathcal{F}}$. En este trabajo se fijó $\lambda = 0{,}25$ para que la frecuencia domine y el % actúe únicamente como desempate.

El ranking consenso $r^\star$ se obtiene ordenando $\mathcal{F}$ por $s(f)$ de forma descendente. La Tabla 4.2 resume dicha lista de referencia, indicando para cada *feature* su frecuencia de aparición en los rankings de cada modelo, su porcentaje medio de importancia relativa y la puntuación final de consenso empleada como relevancia en el cálculo de NDCG.

Feature	**Frecuencia**	**% medio**	**Score consenso**
ZonePercentage	5	4.6892	5.2500
Energy	5	4.6369	5.2459
MaximumProbability	4	4.2435	4.2153
SurfaceVolumeRatio	4	2.6111	4.0882
Imc2	3	2.8488	3.1067
RobustMeanAbsoluteDeviation	3	2.4917	3.0789
Imc1	3	2.3144	3.0651
Correlation	3	1.9531	3.0370
MajorAxisLength	2	2.6824	2.0938
JointEnergy	2	2.5496	2.0834

Tabla 4.2: Ranking consenso $r^\star$ utilizado como lista de referencia para el cálculo de NDCG@10.

Métrica de comparación: NDCG@10.

Para cuantificar la concordancia entre el ranking de cada modelo r_m y el consenso $r^\star$, se empleó la *normalized discounted cumulative gain* truncada en $k = 10$ (NDCG@10), basada en ganancia acumulada con descuento logarítmico y normalización por el ranking ideal [19].

Sea r_m un ranking y sea $\text{rel}(f) \geq 0$ la relevancia asignada a cada variable f (en este trabajo, $\text{rel}(f) = s(f)$). Definimos la ganancia como $g(\text{rel}) = 2^{\text{rel}} - 1$ y el descuento por posición i como $\log_2(i + 1)$. Entonces, para $k = 10$:

$$\text{DCG@10}(r_m) = \sum_{i=1}^{10} \frac{g(\text{rel}(r_m[i]))}{\log_2(i+1)} = \sum_{i=1}^{10} \frac{2^{\text{rel}(r_m[i])} - 1}{\log_2(i+1)}.$$

La normalización se realiza dividiendo por el máximo valor alcanzable (ranking ideal r_{ideal}, obtenido ordenando las variables por rel$(\cdot)$ de forma decreciente):

$$\text{NDCG@10}(r_m) = \frac{\text{DCG@10}(r_m)}{\text{IDCG@10}}, \qquad \text{IDCG@10} = \text{DCG@10}(r_{\text{ideal}}).$$

Valores cercanos a 1 indican alta alineación con el consenso, penalizando más los desacuerdos en posiciones altas que en posiciones bajas.

Conociendo el consenso con el cual se compara, el Cuadro 4.3 resume la NDCG@10 de cada modelo frente a dicho consenso. Se reportan dos variantes: (i) consenso con frecuencia + desempate por % del total (principal), y (ii) consenso solo por frecuencia (control), observándose resultados muy similares.

Modelo	**NDCG@10 (freq+%)**	**NDCG@10 (freq)**
RF	0.9939	0.9948
XGBOOST	0.9397	0.9318
EBM	0.8655	0.8569
SLMVP-SVM	0.7131	0.7154
SVM	0.7006	0.7120

Tabla 4.3: Concordancia de rankings entre modelos mediante NDCG@10 frente al ranking consenso.

Este análisis complementa el basado en medias de importancia, aportando una medida de *robustez estructural* del ranking: RF muestra la mayor alineación con el consenso global, seguido de XGBoost y EBM, mientras que SVM y SLMVP-SVM presentan rankings más divergentes. En conjunto, los resultados sugieren que (i) existe un núcleo de variables recurrentes entre modelos (`ZonePercentage`, `Energy`, `MaximumProbability`, `SurfaceVolumeRatio`), y (ii) la concordancia puede evaluarse de forma independiente a la magnitud absoluta media de importancia.

En resumen, la comparación sugiere que **SLMVP–SVM** ofrece el mejor equilibrio global entre *accuracy/F1* y un *recall* claramente superior en un escenario de alta dimensionalidad y baja muestra, lo que resulta especialmente relevante si se prioriza minimizar falsos negativos; en cambio, la **SVM** maximiza *AUC* y precisión, consistente con una frontera más conservadora (menos falsos positivos) a costa de menor sensibilidad, mientras que **XGBoost** mantiene un rendimiento medio competitivo pero con mayor variabilidad entre particiones y explicaciones más concentradas.

4.0.4 Reentrenamiento del mejor modelo SLMVP–SVM usando solo 8 características (rank = 8)

Una vez identificado el *pipeline* con mejor rendimiento medio (SLMVP–SVM), se planteó un reentrenamiento orientado a parsimonia e interpretabilidad utilizando únicamente un subconjunto compacto de variables. En concreto, se seleccionaron las 8 características con mayor contribución global según las importancias SHAP promediadas entre iteraciones en

los modelos previos: *MaximumProbability, Imc2, Energy, Correlation, ZonePercentage, SurfaceVolumeRatio, Imc1 RobustMeanAbsoluteDeviation.*

La decisión de reentrenar con este subconjunto responde a dos motivaciones complementarias. Desde un punto de vista metodológico, el problema se sitúa en un régimen de alta dimensionalidad y baja muestra ($d = 107$, $n = 62$), donde es fácil que el modelo aprenda patrones espurios dependientes de la partición *train–test*. Reducir el espacio de entrada a variables consistentemente relevantes tiende a disminuir la varianza del estimador, limita la influencia de ruido y redundancia (incluida colinealidad) y, en consecuencia, puede mejorar la generalización del modelo. Por otra parte desde un punto de vista clínico, este paso busca que las decisiones del modelo sean más fácilmente explicables no solo para el radiólogo, sino también para el paciente: un modelo basado en un número pequeño de descriptores radiómicos permite construir explicaciones más transparentes y comunicables (p.ej., asociando la predicción a medidas de energía/intensidad global, heterogeneidad textural y rasgos geométricos), reduciendo la complejidad narrativa necesaria para justificar el resultado.

En coherencia con esta simplificación, el rango del bloque SLMVP se fijó en $r = 8$ ya que es la dimensión máxima de la nueva matriz de proyección.

Manteniendo exactamente el mismo protocolo experimental (15 repeticiones con partición 70–30, ajuste interno por validación cruzada estratificada y evaluación final sobre *test*), este reentrenamiento obtuvo el mejor *accuracy* medio y precisión de entre todos los modelos evaluados.

En promedio, el modelo SLMVP–SVM reentrenado alcanzó:

$$\text{Accuracy} = \mathbf{0{,}7719 \pm 0{,}0736}, \quad \text{F1} = \mathbf{0{,}7396 \pm 0{,}0941}, \quad \text{Recall} = \mathbf{0{,}6741 \pm 0{,}1386},$$

$$\text{Precision} = \mathbf{0{,}8568 \pm 0{,}1182}, \quad \text{AUC} = \mathbf{0{,}8007 \pm 0{,}0714}.$$

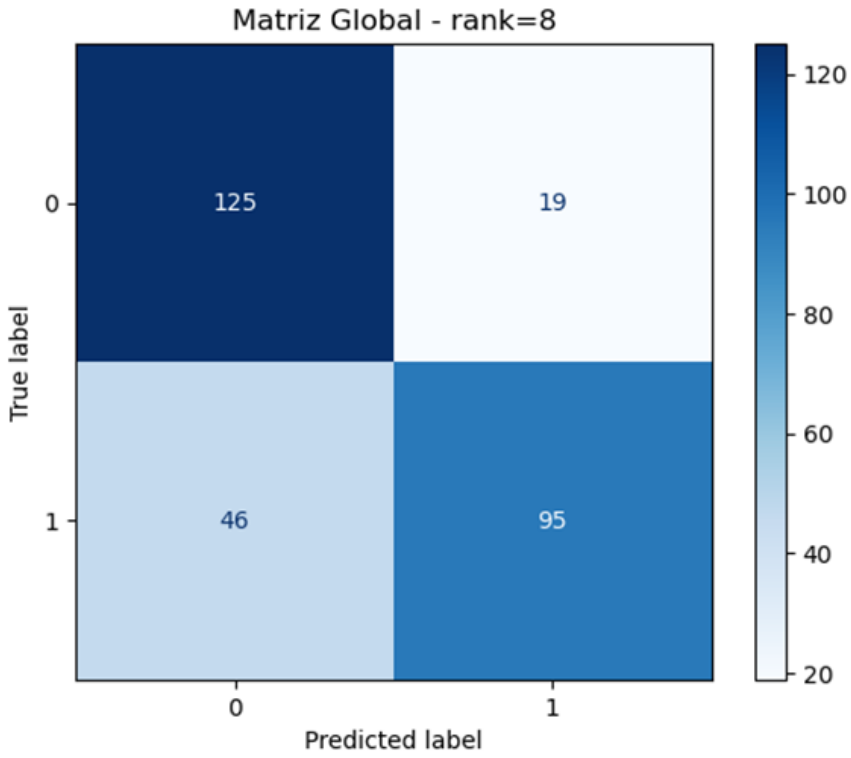


Figura 4.23: Matriz de confusión global del modelo SLMVP–SVM reentrenado con 8 características (rank = 8).

Comparado con la versión SLMVP–SVM original (107 variables, $r = 10$), aunque el reentrenamiento con 8 variables mejora ligeramente la *accuracy* (de 0,7509 a 0,7719) y la *precision* (de 0,7574 a 0,8568), las diferencias globales respecto al SLMVP–SVM completo no son drásticas. Esto es, en gran medida, lo esperable: el *pipeline* ya incorpora una reducción de dimensionalidad supervisada (SLMVP) que actúa como *cuello de botella* y control de capacidad antes de la SVM. Aunque el espacio original tiene $d = 107$ variables, el clasificador opera realmente en un subespacio de dimensión r aprendido para maximizar separabilidad, lo que concentra la señal discriminativa y atenúa la contribución de *features* redundantes o correlacionadas. Por ello, restringir explícitamente la entrada a 8 variables tiende a aportar sobre todo parsimonia e interpretabilidad, mientras que la mejora predictiva adicional queda acotada porque la etapa SLMVP ya estaba diseñada para mitigar el régimen $p \gg n$.

En relación con el resto de modelos estándar, este reentrenamiento no maximiza AUC (la SVM estándar presentó el AUC medio más alto), pero sí ofrece el mejor rendimiento medio en *accuracy* y la mayor precisión, manteniendo además un *recall* superior al de Random Forest, EBM, XGBoost y SVM.

La Figura 4.24 muestra las distribuciones (“campanas”) agregadas de las métricas, y la Figura 4.23 resume el comportamiento global mediante la matriz de confusión agregada. En conjunto, este reentrenamiento mejora el rendimiento global y refuerza la viabilidad clínica del modelo al concentrar la toma de decisión en un conjunto pequeño, estable y más fácilmente justificable de variables radiómicas.

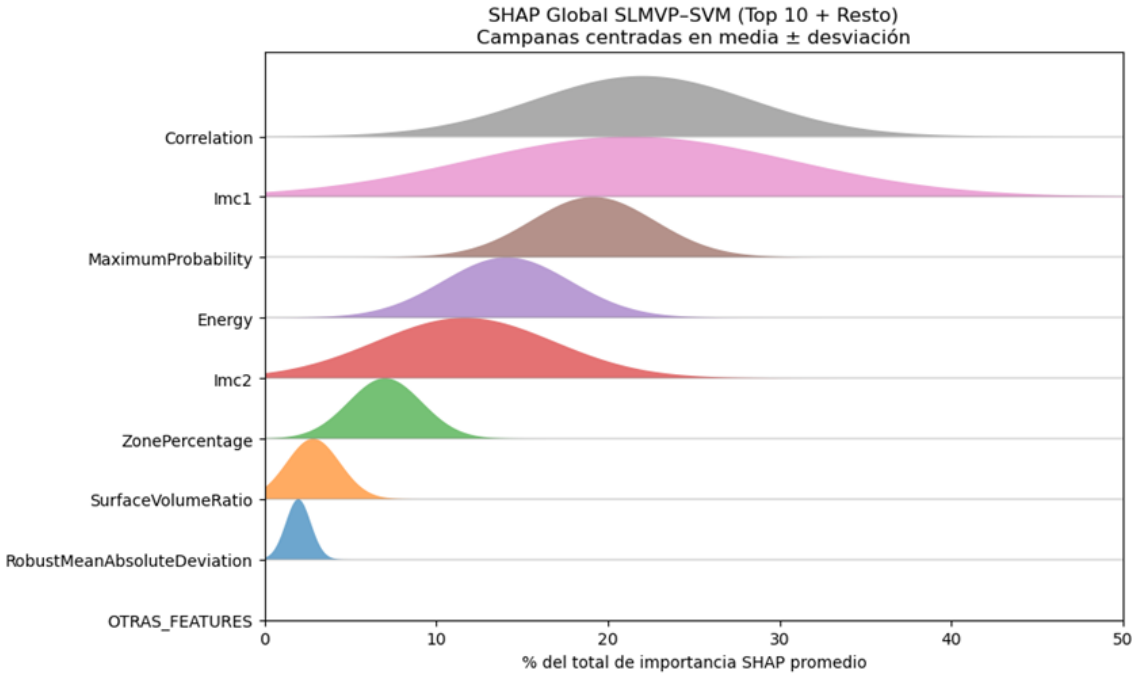


Figura 4.24: Distribuciones agregadas ("campanas") de las métricas del modelo SLMVP–SVM reentrenado usando únicamente las 8 características seleccionadas (rank = 8).

En este modelo, las importancias asignadas por el modelo SLMVP–SVM reentrenado muestran que *Correlation* es la característica dominante (22,01 %), seguida de *Imc1* (21,29 %) y *MaximumProbability* (19,13 %). *Energy* aporta un 14,10 % de la contribución total, mientras que *Imc2* y *ZonePercentage* presentan pesos intermedios (11,66 % y 7,02 %, respectivamente). Finalmente, *SurfaceVolumeRatio* (2,83 %) y *RobustMeanAbsoluteDeviation* (1,96 %) tienen una contribución menor al proceso de decisión.

Capítulo 5

Discusión

Este trabajo aborda un problema clínico —la diferenciación entre absceso cerebral y glioblastoma con realce anular— desde una perspectiva radiómica aplicada a tomografía computarizada (TC), con dos objetivos complementarios: (i) comparar clasificadores representativos de familias habituales en la literatura (ensembles de árboles, modelos aditivos explicables y SVM) y (ii) evaluar la aportación de un *pipeline* propuesto basado en proyección supervisada (SLMVP) seguida de SVM, acompañado de un esquema de explicabilidad diseñado específicamente para este flujo.

5.0.1 Rendimiento, estabilidad y comparación con la bibliografía

En conjunto, los cinco modelos evaluados en el esquema repetido alcanzan *accuracy* medias en el rango 0,72−0,75 y AUC medias en el rango 0,79−0,82. De forma específica, los valores (media ± desviación estándar) fueron: Random Forest $0{,}72 \pm 0{,}05$ (AUC $0{,}81 \pm 0{,}06$), EBM $0{,}72 \pm 0{,}06$ (AUC $0{,}79 \pm 0{,}06$), XGBoost $0{,}73 \pm 0{,}09$ (AUC $0{,}79 \pm 0{,}09$), SVM $0{,}73 \pm 0{,}08$ (AUC $0{,}82 \pm 0{,}06$) y SLMVP–SVM $0{,}75 \pm 0{,}08$ (AUC $0{,}81 \pm 0{,}07$). Esta dispersión no es sorprendente dado el tamaño del conjunto de test en cada repetición: con $n_{\text{test}} = 19$, cada acierto o error modifica el *accuracy* en $\approx 1/19 \simeq 5{,}26\,\%$. Por ello, además del valor medio, resulta crítico interpretar (i) el perfil de error (sensibilidad vs. especificidad) y (ii) la estabilidad (desviación estándar) entre particiones.

Al comparar estos resultados con la bibliografía reciente, se observa que los trabajos centrados en MRI suelen reportar rendimientos superiores, típicamente con AUC en torno a 0,88−0,99 en estudios radiómicos recientes (a menudo usando CE-T1, T2/FLAIR y/o DWI/ADC) y con estrategias de selección de variables y validación interna heterogéneas. En la diferenciación directa entre absceso y glioblastoma, Bo et al. [2] alcanzan AUC = 0,85 (MRI T1/T2, enfoque híbrido HCR+DTL con SVM-RFE), Xiao et al. [44] reportan AUC 0,89−0,99 (MRI CE-T1/T2-FLAIR, con un modelo combinado cercano a 0,97), y Solar et al. [38] informan *accuracy* = 0,85−0,90 (MRI DWI/ADC, kNN/SVM). En este contexto, que en TC los AUC medios queden en 0,79−0,82 es coherente con (i) la menor riqueza de contraste tisular respecto a MRI multiparamétrica y (ii) el hecho de que la literatura con mejores cifras suele apoyarse en secuencias funcionales (p. ej., DWI/ADC) y/o combinaciones de secuencias con alto poder discriminativo microestructural. Aun así, el rendimiento obtenido aquí resulta competitivo dentro del vacío identificado para TC: el presente TFM explora de forma sistemática un *pipeline* radiómico explicable sobre TC, modalidad mucho menos representada y extremadamente útil en la práctica clínica debido

a su velocidad de adquisición de imagen y aplicación a contextos de urgencia médica. Además este trabajo utiliza una cohorte local ($n = 62$) que es comparable o superior a varios trabajos de diferenciación directa (por ejemplo, Solar et al. $n = 40$), aunque menor que otros (Bo et al. $n = 188$, Xiao et al. $n = 118$), y bajo un esquema repetido que permite cuantificar estabilidad, aspecto que no siempre se reporta de forma homogénea en la bibliografía.

Desde el punto de vista clínico, el coste relativo de falsos negativos y falsos positivos depende del circuito asistencial. Etiquetar un absceso como glioblastoma puede retrasar la terapia antibiótica/drenaje, mientras que etiquetar un glioblastoma como un absceso puede demorar decisiones neuroquirúrgicas y oncológicas. En este contexto, resulta apropiado reportar el *F1* como métrica resumen del compromiso *precision–recall*, al integrar ambas magnitudes en un único indicador. La SVM estándar ejemplifica un comportamiento más conservador: alcanza la mayor *precision* media ($0{,}801 \pm 0{,}106$) y la mayor especificidad global (TNR = 0,84), pero con la sensibilidad global más baja (TPR = 0,61; *recall* $0{,}611 \pm 0{,}155$), lo que se traduce en un *F1* inferior ($0{,}678 \pm 0{,}113$). En contraste, SLMVP–SVM desplaza el perfil hacia la detección de la clase positiva: incrementa de forma marcada el *recall* medio a $0{,}745 \pm 0{,}136$ (TPR global = 0,75) manteniendo una especificidad global comparable a Random Forest (TNR = 0,76 en ambos casos) y una *precision* similar ($0{,}757 \pm 0{,}086$). Como resultado, obtiene el mejor *F1* medio ($0{,}743 \pm 0{,}089$). Este cambio es relevante si el objetivo clínico prioriza minimizar falsos negativos de glioblastoma, ya que la mejora respecto a SVM es de $\Delta \approx 0{,}065$ en términos de *F1* (de 0,678 a 0,743).

En este marco encaja el modelo reentrenado (SLMVP–SVM con 8 características y $r = 8$): aunque conserva un F1 comparable ($0{,}74 \pm 0{,}09$) y mejora el *accuracy* medio hasta $0{,}77 \pm 0{,}07$, su *recall* medio baja a $0{,}67 \pm 0{,}14$ frente a $0{,}75 \pm 0{,}14$ del SLMVP–SVM completo. En términos de comparación práctica con la literatura, este resultado ilustra un punto clave: incluso cuando se optimiza y simplifica el modelo (más cercano a los flujos típicos de selección de un subconjunto pequeño de variables), el factor limitante en TC no es solo el clasificador, sino también la información disponible en la modalidad; por ello, la contribución diferencial de este trabajo no se apoya únicamente en maximizar una AUC, sino en (i) caracterizar el compromiso sensibilidad–especificidad de forma transparente y (ii) proponer un *pipeline* explicable y reproducible específicamente sobre TC, donde la evidencia previa es escasa.

5.0.2 Interpretabilidad: por qué cambian los SHAP y qué dicen las *features*

La comparación de *beeswarms* y "campanas" SHAP requiere separar dos factores: (i) el modelo (su hipótesis inductiva) y (ii) el método de explicación (TreeSHAP vs KernelSHAP vs aproximación personalizada). En árboles, TreeSHAP calcula contribuciones aprovechando estructura de decisión; las contribuciones suelen ser más cuantizadas y aparecen agrupaciones verticales (muestras que caen en las mismas hojas o comparten splits relevantes). En KernelSHAP, la contribución depende del muestreo de coaliciones y del background; en alta dimensión y con fuerte correlación, muchas variables reciben contribuciones pequeñas y cercanas al valor base, lo que visualmente produce un *beeswarm* más colapsado como es el caso de SVM. Esta diferencia no implica que un modelo "entienda mejor" que otro, sino que la granularidad de su función y el procedimiento de apagado/encendido cambian la descomposición del efecto.

De entre estos modelos, el que discrimina mejor entre las variables más relevantes es XGBoost. El patrón observado (Las 10 mejores características concentran $\approx 60\,\%$ en XGBoost frente a $\approx 16{-}30\,\%$ en RF/EBM/SVM/SLMVP–SVM) es coherente con dos mecanismos. Primero, el boosting construye árboles secuenciales que corrigen residuales: unas pocas variables pueden dominar si permiten ganancias grandes de pérdida temprano, y el modelo refina alrededor de ellas con interacciones. Segundo, en presencia de variables correlacionadas, los árboles tienden a seleccionar un representante del grupo correlacionado (la primera que ofrece mejor ganancia), y TreeSHAP asigna gran parte del crédito a esa variable elegida, dejando otras correlacionadas con peso menor. Random Forest no discrimina tan fuertemente entre variables porque se basa en bagging y en la selección aleatoria de predictores en cada nodo. Este mecanismo reduce la dependencia de un subconjunto fijo de variables dominantes, ya que distintas réplicas del bosque pueden utilizar predictores diferentes para explicar patrones similares. Como consecuencia, la importancia se reparte entre múltiples variables, especialmente cuando existe colinealidad. En otros modelos (EBM/SVM), el crédito se reparte más, porque varios predictores pueden contribuir de forma marginal similar y la explicación (KernelSHAP) distribuye contribución cuando las variables comparten información.

Limitaciones interpretativas de SHAP en radiómica.

En este régimen, SHAP debe leerse como *atribución condicionada al método*, no como verdad fisiopatológica literal. Con colinealidad, el ranking puede variar: dos variables que codifican la misma información pueden intercambiar posiciones o dividir el peso, y el resultado depende del *background* y de cómo se apagan estas variables (23, 46). Esto es especialmente importante para KernelSHAP y para la implementación personalizada en SLMVP–SVM: aunque el uso de baselines reales muestreados reduce estados irreales, no elimina la dependencia del conjunto de referencia.

Interpretación clínica de las variables más relevantes

Aun con las cautelas anteriores (muestra reducida, posible colinealidad y dependencia de discretización/segmentación), el consenso de variables que emergen en ≥ 3 modelos sugiere qué familias radiómicas capturan diferencias sistemáticas entre absceso y glioblastoma en TC. En el Cuadro 5.1 se sintetiza su grupo y significado, siguiendo las definiciones de PyRadiomics [12]. En conjunto, dominan: (i) textura (GLCM/GLSZM) y (ii) estadísticos de primer orden, con una contribución geométrica (Shape) a través de *SurfaceVolumeRatio*. Esto es conceptualmente consistente con el problema: ambas lesiones pueden compartir realce anular y edema, por lo que el valor añadido suele venir de heterogeneidad interna, organización de intensidades y complejidad del borde/volumen.

Feature	**Grupo**	**Interpretación (PyRadiomics)**
ZonePercentage	GLSZM	Proporción relativa de zonas homogéneas respecto al número total de vóxeles: ZP = $\frac{N_z}{N_v}$. Valores altos suelen asociarse a texturas más "finas" (más zonas pequeñas); valores bajos a zonas más grandes/dominantes. [12]
Energy	First Order	Energía como suma de intensidades al cuadrado en la ROI: Energy = $\sum_{i=1}^{N} x_i^2$. Aumenta con magnitud global de intensidades y/o volumen efectivo tras discretización. [12]
MaximumProbability	GLCM	Máxima probabilidad de co-ocurrencia: MaxProb = $\max_{i,j} P(i,j)$. Valores altos reflejan dominancia de un par de niveles de gris (mayor uniformidad local); valores bajos sugieren mayor heterogeneidad. [12]
Imc2	GLCM	*Informational Measure of Correlation 2* evalúa la dependencia informacional entre intensidades vecinas mediante una transformación de la información mutua. Valores bajos indican independencia entre niveles de gris, mientras que valores altos se asocian a mayor dependencia y complejidad de la textura. [12]
SurfaceVolumeRatio	Shape	Complejidad geométrica (superficie relativa al volumen): SVR = $\frac{\text{SurfaceArea}}{\text{Volume}}$. Tiende a ser mayor en formas más irregulares o menos compactas. [12]
RobustMeanAbsolute Deviation	First Order	Variabilidad robusta de intensidades (MAD robusta), calculada tras recortar extremos (percentiles) para reducir influencia de outliers; captura heterogeneidad de intensidad sin depender de valores extremos. [12]
Imc1	GLCM	*Informational Measure of Correlation 1* también utiliza la información mutua normalizada de la GLCM: valores negativos reflejan mayor dependencia y un patrón de textura más estructurado. [12]
Correlation	GLCM	Dependencia lineal entre niveles de gris vecinos: valores altos indican relación lineal fuerte entre intensidades vecinas; valores bajos, textura menos estructurada linealmente. [12]

Tabla 5.1: Variables recurrentes (≥ 3 entre las 10 mejores): grupo radiómico e interpretación clínica basada en PyRadiomics.

Además del "qué", los *beeswarm* permiten discutir el "hacia dónde": es decir, qué valores concretos empujan la predicción hacia la clase positiva (glioblastoma). Aunque la dirección

exacta puede variar entre modelos por colinealidad y por cómo se discretizan intensidades, en los resultados se observa un patrón muy repetido: **valores altos de *Energy*** y **valores bajos de *ZonePercentage*** y ***MaximumProbability*** tienden a asociarse con glioblastoma (RF, EBM y XGBoost), al que se añaden señales geométricas y de textura específicas en SVM/SLMVP–SVM (p.ej., *LeastAxisLength* alto y *LargeAreaLowGrayLevelEmphasis* bajo en SLMVP–SVM; *SurfaceVolumeRatio* e *Imc2* bajos en la SVM estándar).

1. **Magnitud global de intensidades (*Energy*, valores altos → glioblastoma).** *Energy* aumenta cuando una proporción relevante de la ROI presenta intensidades elevadas y/o cuando el volumen efectivo segmentado es mayor. En TC, esto es compatible con glioblastoma por la coexistencia de tejido tumoral viable, realce irregular, y posibles focos hemorrágicos, que elevan la magnitud global de intensidades al cuadrado. En contraste, un absceso típico presenta una cavidad central hipodensa dominante y un anillo relativamente delgado, lo que tiende a reducir la contribución energética global.

2. **Ausencia de un patrón textural dominante (*MaximumProbability*, valores bajos → glioblastoma).** Los valores bajos de esta característica indican que la probabilidad está repartida entre múltiples pares, lo que es coherente con una textura más compleja y menos repetitiva de los glioblastomas debido a la mezcla espacial de necrosis, tumor viable, edema y realce irregular.

3. **Organización espacial a gran escala (*ZonePercentage*, valores bajos → glioblastoma).** Valores bajosde esta característica sugieren una organización interna basada en regiones extensas de intensidades similares (p. ej., áreas amplias de necrosis o tejido tumoral), separadas por transiciones marcadas, un patrón compatible con la compartimentalización macroscópica típica del glioblastoma. Por el contrario, en los abscesos cerebrales la cavidad y el anillo inflamatorio suelen fragmentarse, tras la discretización de intensidades, en un mayor número de regiones homogéneas pequeñas, lo que incrementa el número de zonas por vóxel y conduce a valores más elevados.

4. **Heterogeneidad robusta de intensidades (*RobustMeanAbsoluteDeviation*).** Valores elevados son consistentes con la coexistencia de múltiples subregiones con intensidades distintas dentro de la ROI, un rasgo frecuente en glioblastoma. Por el contrario, en abscesos, la distribución de intensidades puede estar más concentrada (cavidad hipodensa dominante y anillo relativamente homogéneo), reduciendo esta variabilidad robusta.

5. **Dependencias texturales complejas y estructuradas (*Imc1* bajo, *Imc2* y *Correlation* altos → glioblastoma).** Estas métricas caracterizan la organización espacial de las intensidades más allá de su dispersión. Valores altos de *Imc2* indican una dependencia informacional elevada entre niveles de gris vecinos, reflejando patrones texturales complejos y no aleatorios. En paralelo, valores más negativos de *Imc1* son coherentes con una información mutua alta normalizada, asociada a estructuras internas organizadas pero de elevada complejidad. Finalmente, valores altos de *Correlation* sugieren relaciones lineales locales persistentes entre intensidades adyacentes. En conjunto, este perfil es compatible con el glioblastoma, donde coexisten regiones estructuradas (p. ej., tumor viable, bordes de necrosis, áreas de realce) dentro

de una arquitectura globalmente heterogénea; en contraste, los abscesos tienden a presentar dependencias espaciales más simples y regulares.

6. **Complejidad geométrica y escala (*SurfaceVolumeRatio* alto → glioblastoma).** *SurfaceVolumeRatio* aumenta en formas irregulares, pero también depende del tamaño: a mayor volumen, el cociente puede disminuir incluso si el contorno es irregular. En este estudio, dado que el tamaño puede variar ampliamente entre lesiones y etapas evolutivas (tanto en abscesos como en glioblastomas), es probable que la capacidad discriminativa provenga principalmente de la irregularidad del borde. En este sentido, los glioblastomas tienden a presentar superficies más complejas, con márgenes infiltrativos e irregulares.

7. **Ausencia de cavidades hipodensas extensas (*LargeAreaLowGrayLevelEmphasis*, valores bajos → glioblastoma).** Esta variable aumenta cuando existen áreas grandes y homogéneas de baja intensidad. Valores bajos en glioblastoma son clínicamente plausibles, ya que, aunque pueda haber necrosis, rara vez se presenta como una cavidad amplia y uniformemente hipodensa como en muchos abscesos. La reducción de este énfasis refleja la mayor mezcla de componentes tisulares dentro de la ROI tumoral.

En conjunto, estas características apuntan a dos ejes discriminativos coherentes con lo observado en los *beeswarm*: (i) un eje de **heterogeneidad y complejidad interna**, donde el glioblastoma se caracteriza por una mezcla de intensidades, dependencias texturales no triviales y ausencia de patrones dominantes únicos; y (ii) un eje de **irregularidad geométrica**, asociado a márgenes infiltrativos y organización espacial menos compacta. Frente a ello, los abscesos tienden a mostrar estructuras más simples, cavidades hipodensas bien definidas y patrones texturales más regulares. De esta forma, los resultados sugieren que la separación entre ambas entidades en TC se apoya menos en diferencias de intensidad aisladas y más en la combinación de complejidad estructural, organización espacial y geometría de la lesión, aspectos no evidentes que la radiómica captura de forma complementaria a la evaluación visual convencional.

Valor añadido de SLMVP–SVM en interpretabilidad.

La combinación SLMVP–SVM aporta un valor diferencial en interpretabilidad por un motivo clave: hace que la reducción de dimensionalidad deje de ser una "caja negra" previa al clasificador y pase a formar parte de un *pipeline* explicable de extremo a extremo. En radiómica, la alta dimensionalidad ($d = 107$) y la colinealidad entre descriptores favorecen explicaciones basadas en representantes arbitrarios de grupos correlacionados (típico en árboles) o en repartos muy difusos de contribución (en métodos *model-agnostic*), lo que dificulta construir un relato clínico coherente. En este trabajo, SLMVP introduce una proyección *supervisada* que reorganiza el espacio de variables maximizando la separabilidad guiada por etiquetas [15]; posteriormente, la SVM opera en un subespacio compacto, reduciendo la varianza del modelo frente a particiones pequeñas y favoreciendo explicaciones más estables a nivel global.

Una limitación clásica de las proyecciones supervisadas es que el subespacio latente $z = B^\top \tilde{x}$ no es directamente interpretable en términos de *features* radiómicas originales. El aporte metodológico aquí es resolver esta limitación mediante un SHAP personalizado

que explica el funcional completo

$$f(x) = \mathbb{P}(y = 1 \mid x),$$

evaluando siempre el flujo *escalado* → *SLMVP* → *SVM*, pero asignando contribuciones a las variables originales x_i mediante una aproximación Monte Carlo por permutaciones [46]. Esto permite que la interpretabilidad sea operativa en la práctica clínica:

1. **Trazabilidad clínica sin renunciar a la proyección.** Las explicaciones vuelven al espacio de la radiómica, expresándose en términos familiares para el radiólogo en lugar de interpretar componentes latentes, se identifican familias de descriptores recurrentes (primer orden y texturas GLCM/GLSZM, con apoyo de forma), lo que conecta directamente con hipótesis fisiopatológicas plausibles.

2. **Estabilidad interpretativa en baja muestra.** Con $n_{\text{test}} = 19$ por iteración, el riesgo principal no es solo el rendimiento medio, sino la inestabilidad explicativa entre particiones. Al operar en un subespacio supervisado y evaluar contribuciones usando *baselines* reales muestreados, se evita romper correlaciones y se favorece que el consenso inter-iteración refleje señal robusta y no artefactos de discretización o partición. Esto se manifiesta en la recurrencia de variables clave (p. ej., *Energy*, *ZonePercentage*, *Imc1*, *Correlation*, *SurfaceVolumeRatio*) en relación a los modelos estándar.

Este valor añadido se refuerza con el reentrenamiento usando 8 variables: al concentrar la decisión en un subconjunto reducido, el modelo gana comunicabilidad clínica y permite un relato causal más corto y verificable, aunque con el compromiso habitual precisión–sensibilidad. En conjunto, el aporte de SLMVP–SVM no es solo mejorar métricas, sino convertir la reducción supervisada en una etapa justificable, manteniendo la potencia discriminativa y devolviendo explicaciones estables en variables radiómicas originales, defendibles ante un clínico.

5.0.3 Limitaciones, amenazas a la validez y trabajo futuro

Para que los hallazgos sean interpretables con rigor, es esencial explicitar las amenazas a validez (interna y externa) que condicionan tanto rendimiento como explicabilidad:

1. **Tamaño muestral reducido y test muy pequeño por iteración.** El conjunto total incluye 62 casos y el test por iteración es de 19 casos. Esto amplifica la variabilidad de cualquier métrica: un pequeño número de errores adicionales puede cambiar *accuracy*, *recall* o *precision* varios puntos porcentuales. Aunque se mitiga mediante 15 repeticiones y reporte de media ± desviación estándar, el intervalo de incertidumbre sigue siendo relevante y limita conclusiones finas (p.ej., diferencias pequeñas de AUC entre modelos).

2. **Sesgo de selección del *dataset*.** Los casos fueron seleccionados por criterios radiológicos concretos (lesiones con captación en anillo, exclusiones específicas). Esto es razonable para acotar el problema clínico, pero condiciona el dominio de validez: el modelo aprende sobre una subpoblación de presentaciones y puede no generalizar a lesiones atípicas, a abscesos sin patrón clásico o a tumores con comportamiento sólido u otras variaciones.

3. **Segmentación manual y variabilidad inter/intra-observador.** La delineación manual introduce incertidumbre geométrica y textural: pequeñas diferencias de contorno cambian superficie/volumen y, sobre todo, cambian texturas (GLCM/GLSZM) al modificar el conjunto de vóxeles incluidos. Si no se midió reproducibilidad (p.ej., múltiples segmentaciones), esta fuente de variación debe considerarse una limitación central.

4. **Heterogeneidad de escáner y reconstrucción/interpolación.** Aunque el protocolo se mantuvo homogéneo en la mayoría de casos, existen muestras provenientes de distintos equipos (Toshiba vs Philips/Siemens) y se realizó interpolación a 1 mm desde cortes de 4 mm. Las texturas radiómicas son sensibles a resolución, kernel de reconstrucción y discretización; por tanto, parte de la señal podría reflejar diferencias de adquisición/reconstrucción además de diferencias biológicas.

5. **Ausencia de validación externa.** La falta de un conjunto independiente de otro hospital/protocolo impide evaluar generalización real. En radiómica, donde las distribuciones cambian por escáner y pipeline, esta es probablemente la limitación más importante desde la perspectiva de traslación clínica.

6. **Limitaciones de SHAP (y del SHAP personalizado).** Con *features* correlacionadas, SHAP puede repartir contribución de formas distintas entre variables equivalentes, cambiando rankings sin que cambie el rendimiento predictivo. En KernelSHAP y en el método personalizado, el resultado depende del *background* y de cómo se simula el "apagado" de variables. Por ello, la interpretación correcta es: "qué variables son más útiles para el modelo bajo este esquema de explicación", no "qué variables son causalmente determinantes".

Trabajo futuro.

Para consolidar estos resultados y acercarlos a práctica clínica, los siguientes pasos son naturales: (i) validación externa multicéntrica (idealmente con armonización de *features* entre escáneres), (ii) evaluación explícita de reproducibilidad frente a segmentación (inter/intra-observador o segmentación automática), (iii) incorporación de variables clínicas básicas (edad, marcadores inflamatorios, contexto infeccioso) y (iv) extensión multimodal con RM cuando esté disponible, usando la TC como herramienta rápida inicial y la RM como confirmación/estratificación, en línea con el posicionamiento clínico de ambas técnicas en el estado del arte.

Capítulo 6

Conclusiones

En este Trabajo Fin de Máster se ha abordado el diagnóstico diferencial entre absceso cerebral (AC) y glioblastoma (GBM) con realce anular en tomografía computarizada (TC) mediante un enfoque radiómico y de aprendizaje automático, con énfasis en dos aspectos: (i) comparar modelos representativos de familias habituales (ensembles de árboles, modelos aditivos explicables y SVM) bajo un protocolo de evaluación repetido y (ii) proponer y validar un *pipeline* alternativo basado en proyección supervisada (SLMVP) seguida de SVM, acompañado de un esquema de explicabilidad adaptado a esta arquitectura.

A nivel de datos, se construyó una cohorte balanceada de 62 casos procedentes del Hospital Universitario Ramón y Cajal (31 GBM y 31 AC), con segmentaciones manuales y un preprocesado homogéneo que incluyó interpolación a 1 mm. A partir de cada volumen segmentado se extrajeron 107 características radiómicas (forma, primer orden y varias familias de textura). Este diseño, aunque limitado en tamaño, es relevante por dos motivos: (1) mantiene equilibrio por clase, lo que facilita comparaciones justas entre modelos y (2) explora una modalidad (TC) relativamente poco tratada en la literatura radiómica para este problema, pese a su disponibilidad y rapidez diagnóstica en urgencias.

En cuanto a metodología, se aplicó un esquema robusto para un escenario de muestra pequeña: particiones *train–test* 70–30 repetidas 15 veces, ajuste interno de hiperparámetros con validación cruzada estratificada (maximizando el número de *folds* permitido por el tamaño de *train*) y evaluación final en un conjunto de *test* independiente en cada repetición. Este planteamiento permite estimar no solo el rendimiento medio, sino también su estabilidad (desviación estándar) frente a cambios en la partición, un punto crítico cuando el *test* por iteración es necesariamente pequeño.

Los resultados muestran que los modelos estándar (Random Forest, XGBoost, EBM y SVM) alcanzan rendimientos medios similares, con *accuracy* alrededor de 0.72–0.73 y AUC en torno a 0.79–0.82, lo que confirma que la radiómica en TC contiene señal discriminativa útil incluso sin recurrir a modalidades avanzadas. Sin embargo, el *pipeline* SLMVP–SVM (107 variables, $r = 10$) aporta un mejor equilibrio global entre rendimiento y sensibilidad: obtiene una *accuracy* media de 0.75, un F1 de 0.74 y un *recall* de 0.75, manteniendo además un AUC competitivo. Este patrón es especialmente relevante clínicamente si se prioriza reducir falsos negativos de GBM en un contexto de triaje o apoyo al diagnóstico inicial. Además, se evaluó un reentrenamiento del mejor enfoque orientado a parsimonia e interpretabilidad, restringiendo el modelo a 8 características consensuadas por importancia (rank = 8), el cual alcanzó el mejor rendimiento medio en *accuracy* (0.77) y una precisión

elevada (0.86), a costa de una reducción de sensibilidad (*recall* 0.67). En términos prácticos, el resultado evidencia un compromiso configurable: el modelo completo (más sensible) puede ser preferible cuando se busca "no dejar pasar" GBM, mientras que el modelo compacto (más preciso) puede ser útil cuando se desea minimizar falsos positivos y facilitar explicaciones clínicas más directas.

En explicabilidad, el trabajo refuerza una contribución metodológica clara: se integró un análisis SHAP adaptado a cada familia de modelo (TreeSHAP en árboles y KernelSHAP en modelos no arbóreos) y, de forma diferencial, se desarrolló un SHAP personalizado para el *pipeline* SLMVP–SVM, permitiendo atribuciones en el espacio original de variables pese a la proyección supervisada intermedia. A nivel de consistencia inter-modelo, emergió un núcleo de descriptores recurrentes (p. ej., *Energy*, *ZonePercentage*, *MaximumProbability*, *Imc1/Imc2*, *Correlation*, *SurfaceVolumeRatio*, *RobustMeanAbsoluteDeviation*), lo que sugiere que la discriminación se apoya de forma estable en intensidad global, heterogeneidad textural y complejidad geométrica, un relato compatible con diferencias esperables entre necrosis/infiltración tumoral y cavidad purulenta encapsulada.

Como limitaciones principales, el estudio se basa en una única institución y en segmentación manual, y no incluye validación externa, por lo que la generalización a otros hospitales, protocolos o equipos debe considerarse todavía abierta. También es probable que existan efectos de escáner y de adquisición (aunque el protocolo es mayoritariamente homogéneo) y, como en radiómica en general, la reproducibilidad puede depender de discretización, interpolación y estabilidad de la segmentación. Por ello, los resultados deben interpretarse como una demostración sólida de viabilidad y como una base técnica para estudios confirmatorios.

Como líneas futuras, el siguiente paso natural es validar el *pipeline* en cohortes multicéntricas con armonización de *features* y evaluación prospectiva, incorporar segmentación automática para reducir variabilidad y coste, integrar variables clínicas y/o analíticas (p. ej., marcadores infecciosos) para mejorar el rendimiento en casos límite, y explorar estrategias de calibración/ajuste de umbral según el objetivo clínico (sensibilidad vs especificidad). En conjunto, este TFM muestra que un enfoque radiómico explicable en TC puede apoyar el diagnóstico diferencial AC–GBM con rendimiento competitivo y con trazabilidad interpretativa, sentando una base realista para su futura integración en flujos clínicos asistenciales.

Bibliografía

[1] Ayesha, S., Hanif, M.K., Talib, R., 2020. Overview and comparative study of dimensionality reduction techniques for high dimensional data. Information Fusion URL: https://doi.org/10.1016/j.inffus.2020.01.005, doi:10.1016/j.inffus.2020.01.005.

[2] Bo, L., Zhang, Z., Jiang, Z., Yang, C., Huang, P., Chen, T., Wang, Y., Yu, G., Tan, X., Cheng, Q., Li, D., Liu, Z., 2021. Differentiation of brain abscess from cystic glioma using conventional mri based on deep transfer learning features and hand-crafted radiomics features. Frontiers in Medicine URL: https://doi.org/10.3389/fmed.2021.748144, doi:10.3389/fmed.2021.748144.

[3] Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

[4] Carter, R.M.S., Pretorius, P.M., 2007. The use of ct and mri in the characterization of intracranial mass lesions. Imaging 19, 173–184. doi:10.1259/imaging/64168868.

[5] Cerrone, L., Capuozzo, A., Rocco, B., Nigro, O., Ciancia, G., Santamaria, F., Albino, F., Tortora, G., Ciardo, A., Ciardiello, F., 2022. Radiomics for glioma differential diagnosis: A systematic review based on the radiomics quality score (rqs). Cancers URL: https://doi.org/10.3390/cancers14225712, doi:10.3390/cancers14225712.

[6] Chao, G., Mao, C., Wang, F., Zhao, Y., Luo, Y., 2018. Supervised nonnegative matrix factorization to predict icu mortality risk, in: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). URL: https://doi.org/10.1109/BIBM.2018.8621403, doi:10.1109/BIBM.2018.8621403.

[7] Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. arXiv abs/1603.02754.

[8] Chowdhury, M.Z.I., Turin, T.C., 2020. Precision health through prediction modelling: factors to consider before implementing a prediction model in clinical practice. Journal of Primary Health Care 12, 3–9. URL: https://doi.org/10.1071/HC19087, doi:10.1071/HC19087.

[9] Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297. URL: https://doi.org/10.1007/BF00994018, doi:10.1007/BF00994018.

[10] Desbarats, L.N., Herlidou, S., de Marco, G., Gondry-Jouet, C., Le Gars, D., Deramond, H., Idy-Peretti, I., 2003. Differential mri diagnosis between brain abscesses and necrotic or cystic brain tumors using the apparent diffusion coefficient and normalized diffusion-weighted images. Magnetic Resonance Imaging 21, 645–650. doi:10.1016/

s0730-725x(03)00084-5.

[11] Desprechins, B., Stadnik, K., Koerts, P., Shabana, P., Breucq, C., Osteaux, L., 1999. Use of diffusion-weighted mr imaging in differential diagnosis between intracerebral necrotic tumors and cerebral abscesses. American Journal of Neuroradiology 20, 1252–1257. URL: http://www.ajnr.org/content/20/7/1252.

[12] Documentation, P., . Comprehensive radiomics feature extraction. URL: https://pyradiomics.readthedocs.io/en/latest/features.html. accedido: 15 de Diciembre de 2025.

[13] Fujifilm, a. Synapse3d: Software de visualización y procesamiento de imágenes médicas. URL: https://www.fujifilm.com/ec/es/healthcare/healthcare-it/3d-synapse. accedido: 3 de Septiembre de 2025.

[14] Fujifilm, b. Synapse®pacs. URL: https://www.fujifilm.com/ec/es/healthcare/healthcare-it/it-imaging/pacs. accedido: 3 de Septiembre de 2025.

[15] García-Cuesta, E., Aler, R., del Pozo-Vázquez, D., Galván, I.M., 2023. A combination of supervised dimensionality reduction and learning methods to forecast solar radiation. Applied Intelligence 53, 13053–13066. doi:10.1007/s10489-022-04175-y.

[16] He, X., Niyogi, P., 2003. Locality preserving projections, in: Advances in Neural Information Processing Systems 16 (NIPS 2003), MIT Press, Vancouver, Canada. URL: https://papers.nips.cc/paper/2359-locality-preserving-projections.pdf.

[17] Healthcare, P., . Ct 5300 helical multislice scanner specifications. URL: https://www.philips.es/healthcare/product/HC728285/ct-5300-hc728285-ct-scanner. accedido: 12 de Noviembre de 2025.

[18] Holland, S.M., 2019. Principal Components Analysis (PCA). Technical Report. Department of Geology, University of Georgia. Athens, GA, USA. URL: https://strata.uga.edu/software/pdf/pcaTutorial.pdf.

[19] Järvelin, K., Kekäläinen, J., 2002. Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems 20, 422–446. URL: https://doi.org/10.1145/582415.582418, doi:10.1145/582415.582418.

[20] Kadota, O., Kohno, K., Ohue, S., Kumon, Y., Sakaki, S., Kikuchi, K., Miki, H., 2001. Discrimination of brain abscess and cystic tumor by in vivo proton magnetic resonance spectroscopy. Neurologia medico-chirurgica URL: https://doi.org/10.2176/nmc.41.121, doi:10.2176/nmc.41.121.

[21] Li, X., Chai, Y., Du, M., Lakkaraju, H., Chen, J., Xiong, H., 2023. M4: A unified xai benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities and models, in: Advances in Neural Information Processing Systems.

[22] Liu, Y., Li, T., Fan, Z., Li, Y., Sun, Z., Li, S., Liang, Y., Zhou, C., Zhu, Q., Zhang, H., Liu, X., Wang, L., Wang, Y., 2022. Image-based differentiation of intracranial metastasis from glioblastoma using automated machine learning. Frontiers in Neuroscience URL: https://doi.org/10.3389/fnins.2022.855990, doi:10.3389/fnins.2022.855990.

[23] Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 4765–4774.

[24] Meaney, C., Das, S., Colak, E., Kohandel, M., 2023. Deep learning characterization of brain tumours with diffusion weighted imaging. Journal of Theoretical Biology URL: https://doi.org/10.1016/j.jtbi.2022.111342, doi:10.1016/j.jtbi.2022.111342.

[25] MSD Manual Editorial Board, 2024. Absceso cerebral. *MSD Manual Profesional*. URL: https://www.msdmanuals.com/es/professional/trastornos-neurol%C3%B3gicos/infecciones-cerebrales/absceso-cerebral. accedido el 8 de diciembre de 2025.

[26] Médicos, I.E., . Tomógrafo toshiba aquilion 64 cortes. URL: https://www.isemequiposmedicos.com.mx/catalogo-tomografos/tomografo-toshiba-aquilion-64-cortes/. accedido: 3 de Septiembre de 2025.

[27] National Cancer Institute, 2024. Computed tomography (ct) scans and cancer. *National Cancer Institute*. URL: https://www.cancer.gov/about-cancer/diagnosis-staging/ct-scans-fact-sheet. accedido el 8 de diciembre de 2025.

[28] National Cancer Institute, 2025. Glioblastoma – definition and facts. *National Cancer Institute*. URL: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/glioblastoma. accedido el 8 de diciembre de 2025.

[29] Pomohaci, D., Marciuc, E.A., Dobrovăț, B.I., Popescu, M.R., Istrate, A.C., (Oniciuc), O.M.O., Chirica, S.I., Chirica, C., Haba, D., 2025. Artificial intelligence-based mri segmentation for the differential diagnosis of single brain metastasis and glioblastoma. Diagnostics URL: https://doi.org/10.3390/diagnostics15172248, doi:10.3390/diagnostics15172248.

[30] Priya, S., Liu, Y., Ward, C., Le, N.H., Soni, N., Pillenahalli Maheshwarappa, R., Monga, V., Zhang, H., Sonka, M., Bathla, G., 2021. Machine learning based differentiation of glioblastoma from brain metastasis using mri-derived radiomics. Scientific Reports 11, 10478. URL: https://doi.org/10.1038/s41598-021-90032-w, doi:10.1038/s41598-021-90032-w.

[31] Radiopaedia.org, a. Absceso cerebral | artículo de referencia de radiología. URL: https://radiopaedia.org/articles/cerebral-abscess-1. accedido: 10 de Noviembre de 2025.

[32] Radiopaedia.org, b. Glioblastoma de tipo idh salvaje | artículo de referencia de radiología. URL: https://radiopaedia.org/articles/glioblastoma-idh-wildtype. accedido: 10 de Noviembre de 2025.

[33] Research, M., . Explainable boosting machine (ebm) | interpretml. URL: https://interpret.ml/docs/ebm.html. accedido: 22 de Noviembre de 2025.

[34] Roche, I., . Radiómica. URL: https://www.institutoroche.es/observatorio/radiomica. accedido: 10 de Noviembre de 2025.

[35] Rosales Morales, S., 2024. Absceso cerebral: revisión de tema, fisiopatología, epidemiología, clínica, diagnóstico, microbiología y tratamiento. Revista Electrónica de Portales Médicos 19, 680.

[36] Ruiz-Barrera, M.A., Santamaría-Rodríguez, A.F., Zorro, O.F., 2024. Brain abscess: A narrative review. Neurology Perspectives Médico Fundación Universitaria Juan N. Corpas, Bogotá, Colombia.

[37] Sharma, A., Paliwal, K.K., 2015. Linear discriminant analysis for the small sample size problem: An overview. International Journal of Machine Learning and Cybernetics 6. URL: https://doi.org/10.1007/s13042-013-0226-9, doi:10.1007/s13042-013-0226-9.

[38] Solar, P., Valekova, H., Marcon, P., Mikulka, J., Barak, M., Hendrych, M., Stransky, M., Siruckova, K., Kostial, M., Holikova, K., Brychta, J., Jancalek, R., 2023. Classification of brain lesions using a machine learning approach with cross-sectional adc value dynamics. Scientific Reports URL: https://doi.org/10.1038/s41598-023-38542-7, doi:10.1038/s41598-023-38542-7.

[39] Stupp, R., Mason, W.P., van den Bent, M.J., Weller, M., Fisher, B., et al., 2005. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. New England Journal of Medicine 352, 987–996. doi:10.1056/NEJMoa043330.

[40] Toh, C.H., Wei, K.C., Chang, C.N., Hsu, P.W., Wong, H.F., Ng, S.H., Castillo, M., Lin, C.P., 2012. Differentiation of pyogenic brain abscesses from necrotic glioblastomas with use of susceptibility-weighted imaging. American Journal of Neuroradiology URL: https://doi.org/10.3174/ajnr.A2934, doi:10.3174/ajnr.A2934.

[41] Toh, C.H., Wei, K.C., Ng, S.H., Wan, Y.L., Lin, C.P., Castillo, M., 2011. Differentiation of brain abscesses from necrotic glioblastomas and cystic metastatic brain tumors with diffusion tensor imaging. American Journal of Neuroradiology 32, 1646–1651. doi:10.3174/ajnr.A2581.

[42] Vogelstein, J.T., Bridgeford, E.W., Tang, M., Zheng, D., Douville, C., Burns, R., Maggioni, M., 2021. Supervised dimensionality reduction for big data. Nature Communications URL: https://doi.org/10.1038/s41467-021-23102-2, doi:10.1038/s41467-021-23102-2.

[43] Xia, X., Wu, W., Tan, Q., Gou, Q., 2025. Interpretable machine learning models for differentiating glioblastoma and solitary brain metastasis using radiomics. Academic Radiology URL: https://doi.org/10.1016/j.acra.2025.05.016, doi:10.1016/j.acra.2025.05.016.

[44] Xiao, D., Wang, J., Wang, X., Fu, P., Zhao, H., Yan, P., Jiang, X., 2021. Distinguishing brain abscess from necrotic glioblastoma using mri-based intranodular radiomic features and peritumoral edema/tumor volume ratio. Journal of Integrative Neuroscience URL: https://doi.org/10.31083/j.jin2003066, doi:10.31083/j.jin2003066.

[45] Yi, Z., Long, L., Zeng, Y., Liu, Z., 2021. Current advances and challenges in radiomics of brain tumors. Frontiers in Oncology URL: https://doi.org/10.3389/fonc.2021.732196, doi:10.3389/fonc.2021.732196.

[46] Štrumbelj, E., Kononenko, I., 2011. A general method for visualizing and explaining black-box regression models 6594, 21–30. URL: https://doi.org/10.1007/978-3-642-20267-4_3, doi:10.1007/978-3-642-20267-4_3.