Maestría en Tecnologías de Bases de Datos

Universidad CENFOTEC

Artículo Documento Final de Tesis

Clasificación de tumores cerebrales en imágenes de resonancia magnética mediante redes neuronales convolucionales y transferencia de aprendizaje en el sistema de salud pública costarricense

Harold Garita Matamoros

Abril 2026

Clasificación de tumores cerebrales en imágenes de resonancia magnética mediante redes neuronales convolucionales y transferencia de aprendizaje en el sistema de salud pública costarricense

Harold Garita Matamoros

Resumen:

La detección y clasificación de tumores cerebrales en imágenes de resonancia magnética (MRI) constituye un reto clínico y operativo para los sistemas de salud pública, debido al aumento sostenido en la demanda, la variabilidad interobservador y las limitaciones para procesar grandes volúmenes de estudios. En Costa Rica, la problemática es especialmente crítica en la Caja Costarricense de Seguro Social (CCSS). El Centro Nacional de Imágenes Médicas realiza cerca de 2.970 resonancias al mes, mientras persisten demoras tanto en la realización del estudio como en su interpretación. Según la Auditoría Interna de la CCSS, los pacientes oncológicos esperan en promedio 115 días para acceder a una resonancia, con casos de hasta 648 días, y a diciembre de 2025 se registraban 11.570 estudios pendientes de lectura, con retrasos de hasta 570 días en estudios de cerebro y cuello (Segura, 2026).

Frente a este panorama, se desarrolló y evaluó un sistema de diagnóstico asistido por computadora basado en inteligencia artificial para clasificación multiclase de tumores cerebrales. Se compararon dos arquitecturas representativas, EfficientNet V2 B0 y Vision Transformer (ViT-B16), así como tres estrategias de ensamble basadas en fusión de probabilidades. Los modelos se entrenaron con 4.478 imágenes públicas agrupadas en 44 categorías y se evaluaron sobre un conjunto de prueba independiente (n = 896). EfficientNet V2 B0 alcanzó una exactitud de 92,52% (F1-score ponderado = 0,9250; MCC = 0,9222). Los mejores resultados globales se obtuvieron con el ensamble por promedio ponderado, con 93,30% de exactitud, F1-score ponderado de 0,9327 y MCC de 0,9304.

En suma, los resultados aportan evidencia de que la combinación ponderada de arquitecturas complementarias puede mejorar el desempeño y la robustez en un escenario de 44 clases y desbalance interclase. Estos hallazgos respaldan la aplicabilidad potencial de sistemas de apoyo basados en IA en entornos de salud pública como la CCSS, donde podrían contribuir a estandarizar la clasificación y facilitar la priorización de estudios.

Palabras clave: Redes neuronales convolucionales; transferencia de aprendizaje; clasificación de tumores cerebrales; resonancia magnética; diagnóstico asistido por computadora; aprendizaje profundo

Introducción

La detección de tumores cerebrales en imágenes de resonancia magnética constituye un desafío clínico y técnico importante. En Costa Rica, el volumen de estudios por interpretar ha crecido de manera sostenida cada año. Además, la variabilidad interobservador en la interpretación de las mismas imágenes incrementa la complejidad del proceso. A esto se suman las limitaciones de los sistemas actuales para gestionar grandes volúmenes de datos (Bruner et al., 1997; Dorfner et al., 2025). En este escenario, el tiempo constituye un factor crítico. La prontitud del diagnóstico influye de manera decisiva en el inicio del tratamiento y, en tumores cerebrales, los retrasos prolongados pueden traducirse en progresión de la enfermedad, deterioro neurológico y consecuencias irreversibles, con un impacto directo en la supervivencia del paciente.

En la Caja Costarricense de Seguro Social (CCSS), la magnitud del problema es clara y cuantificable. Aunque el Centro Nacional de Imágenes Médicas realiza alrededor de 2.970 resonancias al mes (Vega, 2025), los retrasos se presentan en dos etapas críticas del flujo diagnóstico. En primer lugar, los pacientes enfrentan demoras para acceder a la realización del estudio. Los pacientes oncológicos esperan en promedio 115 días, con casos de hasta 648 días. En consulta externa, que concentra aproximadamente el 60% de los casos, la espera promedio es de 189 días. En segundo lugar, una vez realizado el examen, también existen demoras para la revisión e interpretación por un especialista. A diciembre de 2025 se registraban 11.570 estudios pendientes de lectura, con retrasos de hasta 570 días en estudios de cerebro y cuello. La Auditoría de la CCSS reportó además la ausencia de una lista organizada de espera y la falta de sistemas automatizados de priorización, lo cual incrementa el riesgo de que los pacientes reciban un diagnóstico tardío (Segura, 2026).

Ante ello, la inteligencia artificial se perfila como una de las alternativas más prometedoras para enfrentar la sobrecarga diagnóstica, ya que permite automatizar la clasificación de imágenes con mayor rapidez y consistencia, y apoyar la toma de decisiones clínicas (Khalighi et al., 2024). Sin embargo, la literatura disponible presenta limitaciones que dificultan implementar esta tecnología a escenarios reales. En particular, una brecha central es que, en gran parte de los estudios disponibles, se entrenan y evalúan modelos con pocas clases y, por ende, con baja

diversidad tumoral. Esto reduce su capacidad de generalización en entornos clínicos, donde la heterogeneidad diagnóstica es frecuente. Además, aunque las estrategias de ensamble pueden aumentar la robustez y la estabilidad al combinar múltiples modelos, su exploración sigue siendo limitada, especialmente en sistemas de salud pública con restricciones operativas (Kilim et al., 2022).

Para abordar estas brechas, este estudio desarrolla y evalúa un sistema de diagnóstico asistido por computadora basado en inteligencia artificial, enmarcado en el sistema de salud público de Costa Rica. La solución se implementa en un entorno de computación en la nube con un flujo de datos estructurado bajo una arquitectura tipo medallón. El sistema aborda un escenario de clasificación multiclase con 44 categorías, lo que permite evaluar su desempeño en un contexto más exigente y representativo que el reportado con frecuencia en la literatura. Además, se compara el rendimiento de dos arquitecturas de aprendizaje profundo, EfficientNet V2 B0 y Vision Transformer (ViT-B16), y se analiza el efecto de distintas estrategias de ensamble sobre la exactitud y la estabilidad de las predicciones, con el objetivo de identificar configuraciones más robustas y viables en condiciones operativas reales.

Metodología

El estudio integra un enfoque cuantitativo, tecnológico y experimental. El componente cuantitativo se centra en la medición y comparación del comportamiento predictivo de los modelos mediante métricas numéricas estandarizadas. El componente tecnológico implica el uso de herramientas de inteligencia artificial y servicios de computación en la nube para la implementación del pipeline de entrenamiento. Por su parte, el enfoque experimental se basa en la evaluación de distintas configuraciones de modelos para identificar cuáles exhiben mejor rendimiento.

Los experimentos se implementaron en Microsoft Azure Machine Learning, utilizando máquinas virtuales con GPU NVIDIA para el entrenamiento de los modelos. El desarrollo se realizó en Python, empleando TensorFlow y Keras para la construcción y entrenamiento de las arquitecturas, y Scikit-learn para el cálculo de métricas de evaluación. El procesamiento y análisis de datos se llevó a cabo mediante NumPy y Pandas, mientras que la visualización de resultados se realizó con Matplotlib.

Fuentes de datos

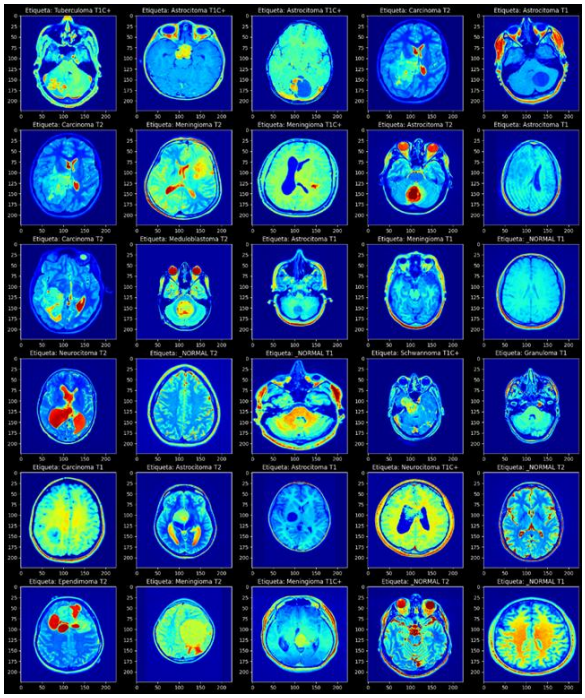
La fundamentación teórica se apoyó en una revisión sistemática de literatura en bases de datos científicas indexadas, incluyendo PubMed, Scopus e IEEE Xplore. La búsqueda inicial identificó 184 estudios potencialmente relevantes, los cuales fueron sometidos a criterios de inclusión y exclusión basados en calidad metodológica y pertinencia temática. Como resultado, se seleccionaron 13 fuentes principales relacionadas con inteligencia artificial aplicada a imágenes médicas, Big Data en salud y clasificación de tumores cerebrales.

Para el entrenamiento y evaluación de los modelos se utilizó un conjunto de datos público de imágenes de resonancia magnética disponible en Kaggle (Feltrin, 2021), compuesto por 4.478 imágenes distribuidas en 44 categorías, correspondientes a 14 tipos de tumores cerebrales y casos normales.

Las imágenes incluyen tres modalidades de adquisición: T1, T2 y T1 con contraste (T1C+). El conjunto de datos se particionó en subconjuntos de entrenamiento (60%), validación (20%) y prueba (20%) mediante muestreo estratificado, garantizando la representación proporcional de cada clase en las tres particiones.

Figura 1

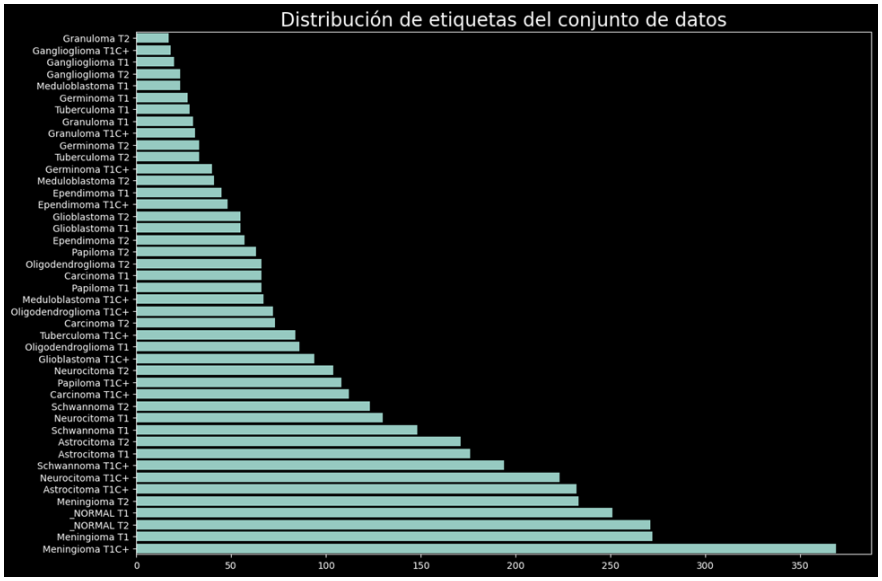
Muestra representativa de imágenes de resonancia magnética del conjunto de datos: 30 imágenes seleccionadas aleatoriamente entre las 44 clases de tumores cerebrales



Nota: Elaboración propia a partir del conjunto de datos Brain Tumor MRI Images 44 Classes (Feltrin, 2021).

Figura 2

Distribución de clases en el dataset de imágenes MRI. Se presentan las 44 categorías de tumores cerebrales ordenadas según su frecuencia de menor a mayor, evidenciando el desbalance de clases presente en el conjunto de datos.



Nota: Elaboración propia a partir del conjunto de datos Brain Tumor MRI Images 44 Classes (Feltrin, 2021).

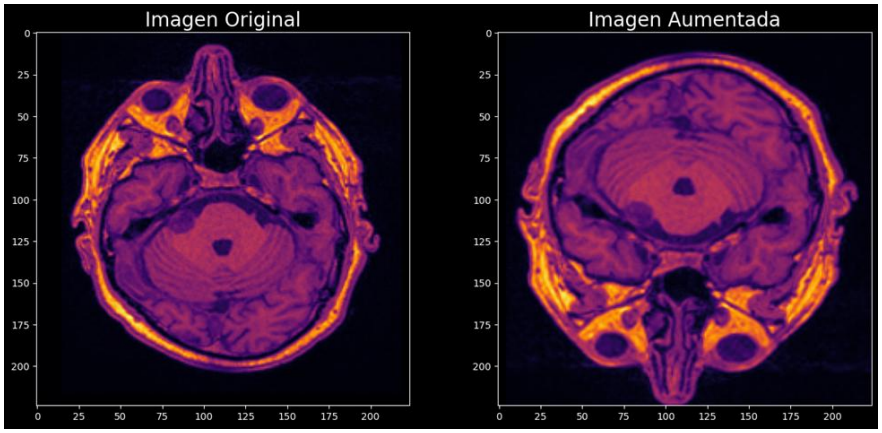
Procedimiento analítico

A partir del conjunto de datos descrito en la sección anterior, se definió un pipeline de preprocesamiento y entrenamiento orientado a la clasificación multiclase de tumores cerebrales. Durante el preprocesamiento, se aplicaron técnicas de aumentación de datos a las imágenes mediante transformaciones estocásticas (p. ej., rotaciones y recortes con zoom), con el objetivo de incrementar la variabilidad del conjunto de entrenamiento y favorecer la estabilidad predictiva del modelo ante variaciones.

Figura 3

Comparación entre imagen MRI original y su versión aumentada. La imagen izquierda muestra la muestra original; la imagen derecha presenta el resultado tras aplicar las técnicas de aumento

de datos: volteo aleatorio horizontal y vertical, y zoom aleatorio de ±10%. Visualización en escala de colores inferno.

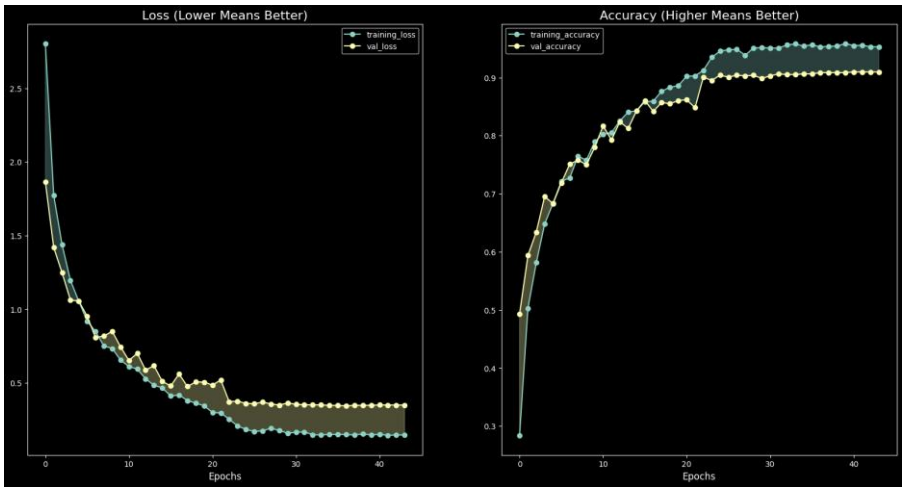


Nota: Elaboración propia a partir del conjunto de datos Brain Tumor MRI Images 44 Classes (Feltrin, 2021), mediante técnicas de aumento de datos implementadas en el pipeline de entrenamiento.

Se implementaron y compararon los modelos EfficientNet V2 B0 y Vision Transformer (ViT-B16), ajustados mediante aprendizaje por transferencia para clasificación multiclase de tumores cerebrales. Como referencia inicial del flujo de entrenamiento se utilizó el notebook de Jansen (2025) en Kaggle, que aborda transfer learning para clasificación de tumores cerebrales. No obstante, el pasar a un problema con 44 clases incrementó la complejidad del espacio de decisión y la dificultad de optimización, por lo que fue necesario recalibrar de manera explícita los hiperparámetros (p. ej., tasas de aprendizaje, regularización y criterios de parada temprana, entre otros ajustes relevantes) y redefinir estrategias de entrenamiento acordes a un escenario multiclase. Estas modificaciones se orientaron, además, a atenuar el efecto del desbalance interclase, con el propósito de reducir el sesgo hacia clases mayoritarias y favorecer la generalización.

Figura 4

Curvas de entrenamiento del modelo EfficientNet V2 B0. El panel izquierdo muestra la evolución de la pérdida (loss) de entrenamiento y validación por época; el panel derecho muestra la exactitud (accuracy) de entrenamiento y validación. Las áreas sombreadas destacan la brecha entre ambas curvas



Nota: Elaboración propia a partir del notebook de Jansen (2025), con modificaciones en los hiperparámetros y el proceso de entrenamiento para adaptarlo a un conjunto de datos desbalanceado y con mayor número de clases.

Resultados

Se comparó de forma sistemática el desempeño de cinco configuraciones: los dos modelos base, EfficientNet V2 B0 y Vision Transformer (ViT-B16), y tres ensambles que combinaron ambos mediante distintas estrategias de fusión de probabilidades (promedio simple, promedio ponderado y media geométrica). El análisis consideró métricas globales y el comportamiento por clase, con el fin de identificar la configuración con mejor equilibrio entre precisión, robustez y capacidad de generalización en un problema de clasificación multiclase.

EfficientNet V2 B0 obtuvo una exactitud del 92,52%, un F1-score ponderado de 0,9250, un MCC de 0,9222 y la mayor exactitud Top-3 del estudio (97,77%), lo que indica que en casi todo el conjunto de prueba la clase verdadera apareció entre las tres predicciones de mayor probabilidad. En síntesis, EfficientNet exhibió el mejor desempeño como modelo individual, con buena capacidad de generalización incluso en clases con bajo soporte, lo cual es coherente con el sesgo inductivo convolucional favorable a patrones locales en MRI.

Por su parte, ViT-B16 alcanzó una exactitud del 82,37%, un F1-score ponderado de 0,8233 y un MCC de 0,8172, con diferencias del orden de 10 puntos porcentuales respecto a EfficientNet en las métricas reportadas. No obstante, mantuvo una exactitud Top-3 del 94,98%, compatible con captura de información discriminativa con menor confianza en la predicción Top-1. Esta brecha es consistente con lo discutido en la literatura fundacional de Vision Transformers en conjuntos de entrenamiento relativamente limitados para la complejidad del arquetipo (en este

estudio, 2.686 imágenes de entrenamiento), donde la ausencia del sesgo de localidad típico de las CNN puede limitar la generalización, especialmente en clases minoritarias.

En cuanto a los ensambles, el promedio simple obtuvo 91,96% de exactitud y un MCC de 0,9166, inferior al EfficientNet individual, lo cual se interpreta por la asimetría de rendimiento entre modelos base: cuando el modelo más débil asigna alta probabilidad a una clase incorrecta, el promediado equitativo puede desplazar la predicción combinada lejos de la clase correcta sugerida por EfficientNet. La media geométrica alcanzó 92,63% de exactitud y un MCC de 0,9235, superando levemente a EfficientNet en exactitud global, pero sin alcanzar al ensamble ponderado. Finalmente, el ensamble por promedio ponderado (pesos 0,6 para EfficientNet y 0,4 para ViT-B16) obtuvo el mejor resultado global entre las cinco configuraciones, con exactitud del 93,30% (mejora de 0,78 puntos porcentuales respecto a EfficientNet), F1-score ponderado de 0,9327 y MCC de 0,9304, además de mejoras en precisión y recall ponderados, lo que apunta a que la contribución de ViT introduce señal complementaria que corrige parte de los errores sistemáticos de la CNN sin anular su mayor precisión individual.

Figura 5

Comparación de rendimiento entre modelos de clasificación

Modelo	Accuracy	Top-3 Accuracy	Precisión	Recall	F1-score	MCC
EfficientNet V2 B0	0.9252	0.9777	0.9311	0.9252	0.925	0.9222
Vision Transformer (ViT-B16)	0.8237	0.9498	0.8482	0.8237	0.8233	0.8172
Ensamble (promedio simple)	0.9196	0.9754	0.9294	0.9196	0.9194	0.9166
Ensamble (promedio ponderado)	0.933	0.9754	0.939	0.933	0.9327	0.9304
Ensamble (media geométrica)	0.9263	0.9732	0.9342	0.9263	0.9261	0.9235

Nota: Elaboración propia a partir de los resultados obtenidos en el experimento.

Al revisar los resultados por cada tipo de tumor, se encontró que los tipos con más imágenes de entrenamiento tuvieron resultados más altos y estables, con precisión y recall por encima del 90%. Los tipos con pocas imágenes mostraron mayor variabilidad y en algunos casos el F1-score bajó de forma notable. Este patrón es coherente con el desbalance de clases visible en la Figura 2, y confirma que tener suficientes datos de entrenamiento es fundamental para que el modelo clasifique bien cada categoría. En particular, debe considerarse que las métricas por clase en categorías con muy pocas observaciones en prueba pueden ser altamente variables y

deben interpretarse con cautela dentro de un entorno experimental controlado, especialmente cuando el soporte es insuficiente para estimaciones estables.

Figura 6

Reporte de clasificación del ensamble por promedio ponderado evaluado sobre el conjunto de prueba (n = 896 imágenes). Se presentan precisión, recall y F1-score por clase, junto con los promedios macro y ponderado. Accuracy global: 0.93.

Clase	Precision	Recall	F1-score	Support
Astrocitoma T1	0.97	0.94	0.96	36
Astrocitoma T1C+	0.95	0.91	0.93	44
Astrocitoma T2	0.95	0.72	0.82	29
Carcinoma T1	1.00	0.87	0.93	15
Carcinoma T1C+	1.00	0.86	0.93	22
Carcinoma T2	1.00	1.00	1.00	18
Ependimoma T1	1.00	1.00	1.00	4
Ependimoma T1C+	1.00	0.86	0.92	7
Ependimoma T2	0.87	0.87	0.87	15
Ganglioglioma T1	1.00	1.00	1.00	2
Ganglioglioma T1C+	1.00	1.00	1.00	7
Ganglioglioma T2	0.50	0.33	0.40	3
Germinoma T1	1.00	1.00	1.00	2
Germinoma T1C+	0.78	1.00	0.88	7
Germinoma T2	0.67	1.00	0.80	10
Glioblastoma T1	1.00	0.83	0.91	6
Glioblastoma T1C+	0.94	0.89	0.92	19
Glioblastoma T2	1.00	0.73	0.84	11
Granuloma T1	1.00	1.00	1.00	9
Granuloma T1C+	1.00	0.89	0.94	9
Granuloma T2	1.00	0.50	0.67	4
Meduloblastoma T1	0.83	1.00	0.91	5
Meduloblastoma T1C+	1.00	0.94	0.97	16
Meduloblastoma T2	1.00	0.83	0.91	6
Meningioma T1	0.92	0.95	0.94	64
Meningioma T1C+	0.95	0.92	0.93	76
Meningioma T2	0.91	0.87	0.89	46
Neurocitoma T1	0.93	1.00	0.96	26
Neurocitoma T1C+	0.95	1.00	0.97	38
Neurocitoma T2	0.92	0.96	0.94	23
Oligodendroglioma T1	1.00	1.00	1.00	12
Oligodendroglioma T1C+	0.95	0.95	0.95	20
Oligodendroglioma T2	1.00	0.89	0.94	9
Papiloma T1	1.00	0.92	0.96	12
Papiloma T1C+	1.00	0.95	0.97	19
Papiloma T2	1.00	0.92	0.96	13
Schwannoma T1	1.00	1.00	1.00	29
Schwannoma T1C+	0.92	1.00	0.96	46
Schwannoma T2	0.90	0.90	0.90	30
Tuberculoma T1	1.00	1.00	1.00	4
Tuberculoma T1C+	1.00	0.89	0.94	9
Tuberculoma T2	1.00	1.00	1.00	8
_NORMAL T1	0.97	1.00	0.98	57
_NORMAL T2	0.77	1.00	0.87	49
accuracy			**0.93**	**896**
macro avg	**0.94**	**0.91**	**0.92**	**896**
weighted avg	**0.94**	**0.93**	**0.93**	**896**

Nota: Elaboración propia a partir de los resultados obtenidos en el experimento.

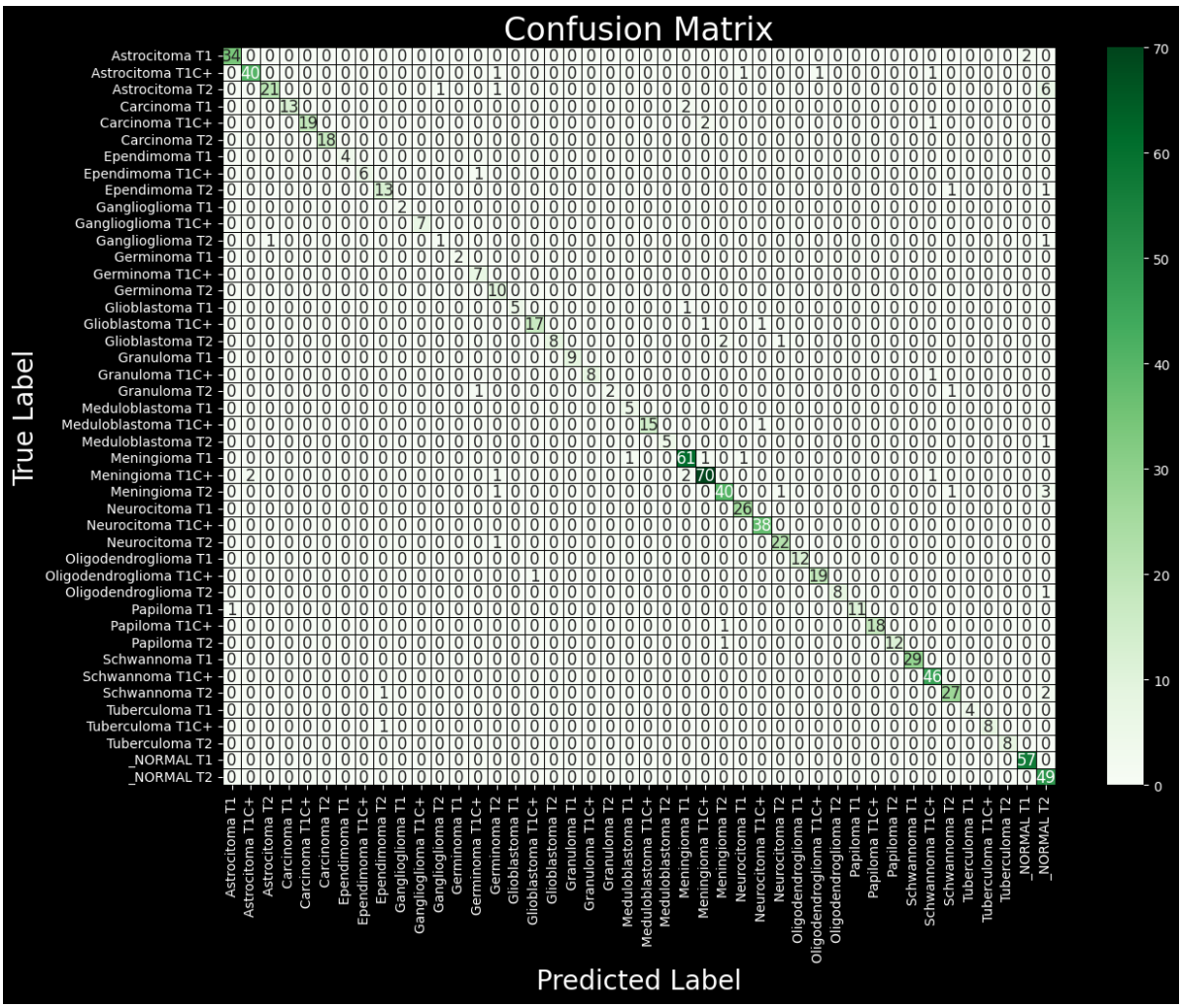
El reporte de clasificación por clase del ensamble por promedio ponderado (Figura 6) permite evaluar el comportamiento a nivel de cada una de las 44 categorías e identificar dónde el modelo generaliza con mayor dificultad. En términos agregados, el ensamble alcanzó accuracy

= 0,93, con F1-score macro = 0,92 y F1-score ponderado = 0,93 (n = 896), lo cual indica un desempeño sólido tanto en clases frecuentes como, en menor medida, en clases minoritarias. Este modelo presentó el mayor número de categorías con F1 = 1,00, incluyendo Carcinoma T2, Ependimoma T1, Ganglioglioma T1, Ganglioglioma T1C+, Germinoma T1, Granuloma T1, Granuloma T1C+, Oligodendroglioma T1, Schwannoma T1, Tuberculoma T1 y Tuberculoma T2.

Paralelamente, se observaron mejoras puntuales respecto a EfficientNet V2 B0 en clases donde el modelo convolucional presentaba mayor dificultad como, por ejemplo, Oligodendroglioma T2 (F1 de 0,80 a 0,94), Ependimoma T2 (de 0,80 a 0,87) y Meduloblastoma T2 (de 0,80 a 0,91). Estas mejoras apoyan la hipótesis de sinergia entre arquitecturas, ya que el ViT-B16 logra clasificar correctamente ciertas muestras que EfficientNet tiende a confundir, y la ponderación 0,6/0,4 permite incorporar esa señal sin comprometer el desempeño global. No obstante, las clases con soporte extremadamente bajo mantuvieron alta variabilidad y menor confiabilidad estadística. Por ejemplo, Ganglioglioma T2 continuó siendo una de las categorías más difíciles (F1 = 0,40, support = 3), lo que refleja la limitación inherente a la escasez de muestras.

Figura 7

Matriz de confusión del ensamble por promedio ponderado.



Nota: Elaboración propia a partir de los resultados obtenidos en el experimento.

La matriz de confusión del ensamble por promedio ponderado evidencia una diagonal principal dominante, lo que refleja el comportamiento agregado del modelo. Los errores fuera de la diagonal se concentran en dos patrones principales. Primero, se observan confusiones entre variantes de secuencia del mismo tipo tumoral (p. ej., T1 clasificado como T1C+), lo cual indica que, aunque el modelo identifica la lesión, la distinción fina por modalidad de adquisición constituye un reto inherente a la granularidad del problema. Segundo, persisten errores en clases con baja representación, donde la generalización es más limitada independientemente de la arquitectura utilizada.

Un patrón clínicamente relevante que se observa de forma consistente es la confusión de algunas muestras de Astrocitoma T2 con NORMAL T2, lo cual implica un riesgo potencial de falso negativo en la secuencia T2. Este hallazgo refuerza que el uso del sistema debe considerarse como apoyo y no como sustituto del juicio clínico, y que la interpretación ideal debe integrar múltiples secuencias (especialmente cuando el realce con contraste en T1C+ aporta señales discriminativas que no siempre están presentes en T2).

Considerando el conjunto del experimento, el problema evaluado implica 44 clases definidas por combinación de tipo de lesión y modalidad de MRI (T1, T1C+ y T2), una granularidad notablemente mayor que la de muchos trabajos de referencia que suelen reportar entre 3 y 5 categorías. En ese contexto, una exactitud global del 93,30% con MCC de 0,9304 constituye un resultado técnicamente sólido. La exactitud Top-3 del ensamble ponderado (97,54%) es particularmente relevante desde una perspectiva de apoyo clínico, ya que reduce el espacio de diagnósticos diferenciales a las tres hipótesis más probables incluso cuando la predicción principal no coincide con la clase verdadera.

Discusión

Los resultados obtenidos confirman que los modelos de aprendizaje profundo pueden alcanzar un elevado rendimiento en la clasificación de tumores cerebrales en imágenes de resonancia magnética bajo esquemas de transferencia de aprendizaje. En particular, EfficientNet V2 B0 superó a Vision Transformer (ViT-B16) en aproximadamente 10 puntos porcentuales en el coeficiente de correlación de Matthews (MCC), lo que evidencia una mayor consistencia en sus predicciones. Este hallazgo se alinea con estudios previos que reportan un mejor desempeño de arquitecturas convolucionales en escenarios con conjuntos de datos limitados, debido a su sesgo inductivo hacia la captura de patrones locales en imágenes médicas (Dorfner et al., 2025; Sadr et al., 2025).

El menor rendimiento de ViT-B16 puede interpretarse en relación con la sensibilidad de los Transformers al tamaño del conjunto de entrenamiento y a esquemas de preentrenamiento a gran escala, ampliamente discutidos en la literatura fundacional de Vision Transformers (Dosovitskiy et al., 2021). En el presente estudio, el tamaño del conjunto de entrenamiento resulta limitado frente a la complejidad de la arquitectura, lo que podría restringir su capacidad de generalización. Sin embargo, el modelo mantuvo una exactitud Top-3 alta, lo cual sugiere que captura información discriminativa relevante, aunque con menor confianza en la predicción Top-1.

Uno de los aportes más relevantes de este estudio es la evidencia empírica de que la combinación de arquitecturas complementarias puede incrementar la calidad predictiva en escenarios multiclase y desbalanceados. El ensamble por promedio ponderado superó tanto a ViT-B16 como a EfficientNet V2 B0 en la mayoría de las métricas evaluadas, en línea con literatura que destaca la capacidad de los ensambles para reducir varianza y atenuar errores sistemáticos de modelos individuales (Khalighi et al., 2024). En este caso, la integración de un modelo basado en atención (Vaswani et al., 2017) aporta información complementaria al sesgo local de las CNN, lo que mejora la robustez del sistema sin comprometer la precisión global.

Más allá del desempeño cuantitativo, los resultados tienen implicaciones relevantes para la práctica clínica. En línea con Chen et al. (2024), la integración de sistemas de inteligencia artificial en el flujo de trabajo médico no busca reemplazar al especialista, sino apoyar la toma de decisiones mediante la priorización y el filtrado inicial de casos. En el contexto del sistema de salud costarricense, donde existen retrasos significativos en la interpretación de estudios de resonancia magnética (Segura, 2026), un sistema con una exactitud superior al 93% podría contribuir significativamente a optimizar la asignación de recursos y reducir los tiempos de diagnóstico.

Por otra parte, el uso de infraestructura en la nube y pipelines automatizados responde a limitaciones operativas previamente documentadas, como la ausencia de sistemas estructurados de gestión y priorización de estudios. La automatización del proceso no solo mejora la eficiencia, sino que también contribuye a reducir la variabilidad interobservador, un fenómeno ampliamente reportado en la literatura clínica (Bruner et al., 1997). En la misma línea, los resultados respaldan el potencial de estos sistemas para estandarizar procesos diagnósticos en entornos de alta demanda.

Adicionalmente, la implementación de soluciones basadas en inteligencia artificial puede contribuir a reducir desigualdades en el acceso a servicios especializados. Tal como ha sido señalado en estudios recientes sobre salud digital, la disponibilidad de herramientas diagnósticas en la nube permite extender capacidades clínicas a regiones con menor acceso a especialistas (Khalighi et al., 2024). En el caso de Costa Rica, esto podría traducirse en una mejora significativa en la equidad del sistema de salud, particularmente en zonas rurales.

No obstante, es importante reconocer las limitaciones del estudio en el contexto de un posible despliegue clínico. Aunque los modelos alcanzaron métricas de elevado rendimiento, la literatura enfatiza la necesidad de validación externa y evaluación prospectiva antes de su

implementación en entornos reales. Desde esta perspectiva, el uso de datos públicos y la ausencia de validación clínica directa limitan la generalización de los resultados. Del mismo modo, el desbalance entre clases afectó el desempeño en categorías con menor representación, lo cual es consistente con lo reportado en estudios previos sobre clasificación multiclase en imágenes médicas.

Finalmente, futuras líneas de investigación deberían orientarse a la construcción de conjuntos de datos institucionales validados clínicamente, la exploración de arquitecturas híbridas que integren de manera más efectiva CNN y Transformers, y la evaluación del impacto de estos sistemas en entornos hospitalarios reales. En particular, resulta relevante complementar las métricas de desempeño con indicadores operativos, como tiempos de diagnóstico y carga de trabajo, tal como se ha propuesto en estudios recientes sobre implementación de inteligencia artificial en salud.

Conclusión

Este estudio demuestra que los sistemas de diagnóstico asistido por computadora basados en aprendizaje profundo pueden alcanzar un rendimiento elevado en la clasificación multiclase de tumores cerebrales en imágenes de resonancia magnética, incluso en escenarios de alta granularidad y desbalance interclase. En particular, el ensamble por promedio ponderado logró el mejor rendimiento global, evidenciando que la combinación de arquitecturas complementarias constituye una estrategia efectiva para mejorar la robustez y la consistencia de las predicciones.

Los resultados confirman que la selección arquitectónica es un factor determinante en este dominio. Las arquitecturas convolucionales mostraron un desempeño superior en condiciones de datos limitados, mientras que la incorporación de modelos basados en atención aportó información complementaria que permitió mejorar el rendimiento global mediante estrategias de ensamble.

Desde una perspectiva aplicada, estos hallazgos son altamente relevantes para el sistema de salud público costarricense, donde persisten desafíos en el acceso y la interpretación oportuna de estudios de resonancia magnética. En este marco, la integración de herramientas de inteligencia artificial puede contribuir a optimizar la priorización de casos, reducir la carga operativa y mejorar la eficiencia diagnóstica.

Por último, el estudio amplía la evidencia existente al evaluar modelos y ensambles en un escenario multiclase más complejo que el abordado comúnmente en la literatura, y establece una base metodológica para futuras investigaciones. En particular, se identifican como líneas prioritarias la validación externa con datos clínicos institucionales, el abordaje del desbalance de clases y la evaluación prospectiva del impacto de estos sistemas en entornos hospitalarios reales.

Referencias

Bruner, J. M., Inouye, L., Fuller, G. N., & Langford, L. A. (1997). Diagnostic discrepancies and their clinical impact in a neuropathology referral practice. *Cancer. 79*(4), 796–803. https://doi.org/10.1002/(SICI)1097-0142(19970215)79:4<796::AID-CNCR17>3.0.CO;2-V

Chen, M., Wang, Y., Wang, Q., Shi, J., Wang, H., Ye, Z., Xue, P., & Qiao, Y. (2024). Impact of human and artificial intelligence collaboration on workload reduction in medical image interpretation. *npj Digital Medicine, 7*, 349. https://doi.org/10.1038/s41746-024-01328-w

Dorfner, F. J., Patel, J. B., Kalpathy-Cramer, J., Gerstner, E. R., & Bridge, C. P. (2025). A review of deep learning for brain tumor analysis in MRI. *npj Precision Oncology, 9*, Article 2. https://doi.org/10.1038/s41698-024-00789-2

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR).* https://arxiv.org/abs/2010.11929

Feltrin, F. (2021). Brain tumor MRI images (44 classes). Kaggle.

https://www.kaggle.com/datasets/fernando2rad/brain-tumor-mri-images-44c/data

Jansen, M. (2025). Transfer Learning | Brain Tumor Classification. Kaggle. https://www.kaggle.com/code/matthewjansen/transfer-learning-brain-tumor-classification

Khalighi, S., Reddy, K., Midya, A., Pandav, K. B., Madabhushi, A., & Abedalthagafi, M. (2024). Artificial intelligence in neuro-oncology: advances and challenges in brain tumor diagnosis, prognosis, and precision treatment. *npj Precision Oncology, 8*, Article 80. https://doi.org/10.1038/s41698-024-00575-0

Kilim, O., Olar, A., Joó, T., Palicz, T., Pollner, P., Csabai, I., … et. al. (2022). Physical imaging parameter variation drives domain shift. *Scientific Reports, 12*, Article 21302. https://doi.org/10.1038/s41598-022-23990-4

Sadr, H., Nazari, M., Yousefzadeh-Chabok, S., Emami, H., Rabiei, R., & Ashraf, A. (2025). Enhancing brain tumor classification in MRI images. *Image and Vision Computing*. https://doi.org/10.1016/j.imavis.2025.105555

Segura, A. (2026). Pacientes con cáncer esperan más de 600 días por una resonancia magnética en la CCSS. CRHoy. https://crhoy.com/nacionales/pacientes-con-cancer-esperan-mas-de-600-dias-por-una-resonancia-magnetica-en-la-ccss/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*. https://arxiv.org/abs/1706.03762

Vega, Y. (2025, 21 agosto). *Centro Nacional de Imágenes Médicas realiza cada mes 2 970 resonancias magnéticas.* https://www.ccss.sa.cr/noticias/noticia?v=012136000152