Clasificación de lesiones mamarias utilizando aprendizaje profundo en imágenes de resonancia magnética.

Erika Alexandra García Barrios
ORCID ID (0009-0005-6751-4174)

Trabajo de grado presentado para optar al título de Ingeniera de Software

Asesores:

Rubén Darío Fonnegra Tarazona
Ph.D en Ingeniería

Mateo Rico García
M.Sc en automatización y control industrial

INSTITUCIÓN UNIVERSITARIA PASCUAL BRAVO
FACULTAD DE INGENIERÍA
INGENIERÍA DE SOFTWARE
MEDELLÍN
2026

Dedicatoria y agradecimientos.

Primeramente, quiero dedicar este trabajo a mi familia, especialmente a mi madre adoptiva, Lucila López, quién desde niña siempre me inculcó la importancia de la educación y a quién le debo haber logrado mantenerme con vida durante mis primeros años. También agradezco profundamente su apoyo emocional durante mi adultez y el hecho de creer fervientemente en mí, incluso en los peores momentos.

A mi hermano Carlos Andrés García Barrios, quién fue clave para mi regreso a la universidad, brindándome su amor, apoyo económico, confianza, ánimo y soporte emocional para continuar adelante con lo que me faltaba de pregrado, gracias por siempre estar ahí para mí y por ser mi confidente.

A mi sobrino Andrés Felipe García Moya, quién sin saberlo, ha transformado mi manera de pensar, acercándome más a mi familia y despertando en mí un deseo ferviente de mejorar como persona, cómo profesional, y cómo tía. Me alegra profundamente saber qué desde sus primeros años de vida está aprendiendo programación y siguiendo mis pasos.

A mis profesores, qué durante toda la carrera fueron comprensivos con las múltiples situaciones personales y académicas que se fueron presentando y algunos, me dieron la mano. Nunca me desmotivaron para desistir de la carrera, por el contrario, me brindaron apoyo, comprensión y motivación constante para continuar con mi proceso académico. Gracias a

muchos de ellos, encontré inspiración para seguir creciendo profesionalmente y fortalecer mi deseo de continuar en el camino de la investigación.

También quiero agradecer a mis asesores, Rubén Fonnegra y Mateo Rico García, quienes aportaron orientación, recomendaciones académicas y acompañamiento durante el desarrollo de este proyecto de investigación.

Por último, y no menos importante, agradezco a mi compañera de estudio, mi gata Athena, quien durante los últimos años fue un apoyo importante para mi bienestar emocional. Estuvo presente en incontables noches de estudio y trabajo, acompañándome con tranquilidad, amor y cariño en muchos de los momentos más difíciles de este proceso académico, ayudándome a sobrellevar la ansiedad, el cansancio y los desvelos que hicieron parte de este camino.

Contenido

Pág.

Tabla de contenido

Resumen 10
Introducción 15
1. Planteamiento del problema 16
1.1 Descripción 16
1.2 Limitaciones de la mamografía 17
1.3 Complejidad de la resonancia magnética 18
1.4 Barreras de acceso y demoras diagnósticas en Colombia 19
1.4.1 Diagnóstico tardío 19
1.4.2 Demoras documentadas 19
1.4.3 El aprendizaje profundo como base para futuras soluciones de ingeniería 20
1.5 Formulación 22
2. Justificación 23
2.1. Justificación social y clínica 23
2.1.1 Problema en los tiempos de diagnóstico: 23
2.1.2 Necesidad de un diagnóstico consistente: 24
2.2. Relevancia tecnológica y científica 25
2.2.1 Representación multicanal de imágenes DCE-MRI: 25
2.2.2 Evaluación experimental de arquitecturas de aprendizaje profundo 26
2.2.3 Transferencia de tecnología: 26
2.3. Relevancia metodológica 27
3. Objetivos 28
3.1 Objetivo general 28
3.2 Objetivos específicos 28
4. Marco teórico 29
4.1. Antecedentes de la clasificación asistida por computadora en imagenología 29
4.1.1. Evolución de los sistemas CAD en mama 29
4.1.2. El Paradigma del aprendizaje profundo en Radiología 30
4.2. Bases clínicas, radiológicas y principios del diagnóstico 30
4.2.1. Principios de la resonancia magnética dinámica con contraste 30
4.2.2. Naturaleza de los datos 3D y dinámica temporal en DCE-MRI 31
4.2.3 Curva persistente (Tipo 1: benigna) : 31

4.2.4 Curva de meseta (tipo 2: sospechosa): 32
4.2.5 Curva de lavado (Tipo 3): 33
4.3. Sistema de reporte de datos de imagen mamaria 34
4.4. Fundamentos de ingeniería y enfoque teórico (aprendizaje profundo) 35
4.4.1. Teoría de las redes neuronales convolucionales (CNN) 35
4.4.2. Transformers visuales y mecanismos de atención 38
4.4.3 Visión transformers (ViT) para la Imagenología 39
4.4.4 Atención global y representaciones multicanal 41
4.5 Métricas de evaluación de rendimiento 41
4.5.1 Curva ROC y área bajo la curva (AUC) 41
4.5.2 Matriz de confusión y métricas clínicas 41
5. Metodología 43
5.1 Dataset y preparación de datos 43
5.1.1 Descripción del dataset BreastDM 43
5.1.2 Organización de imágenes ROI 44
5.1.3 Representación multicanal DCE-MRI 46
5.1.4 Normalización y preprocesamiento 47
5.1.5 Estrategias de aumento de datos 48
5.2 Tipo de proyecto 49
5.3 Método de investigación 50
5.3.1 Procedimiento metodológico: 50
5.3.2 Fase de análisis y preprocesamiento: 51
5.3.3 Fase de implementación y experimentación: 51
5.3.4 Fase de evaluación comparativa: 51
5.3.5 Origen de los datos y tipología de fuentes 52
5.3.5.1 Fuentes primarias. 52
5.3.5.2 Fuentes secundarias: 52
5.4 Arquitecturas implementadas 53
5.4.1 Transfer learning y Fine-tuning 54
5.4.2 Configuración de entrenamiento 56
5.4.3 Focal loss y optimización clínica 57
5.4.4 Calibración probabilística y ajuste de umbrales 58
5.4.5 Entorno experimental y trazabilidad de experimentos 59
5.5 Desarrollo experimental 60
5.5.1 Métricas utilizadas 61
5.5.2 Curvas ROC y Precision - Recall 61
5.5.3 Matrices de confusión 61
5.5.4 Curvas de calibración 62

5.5.5 Interpretabilidad mediante GradCAM 63
6. Resultados y discusión 64
6. 1. Resultados de entrenamiento y convergencia 64
6. 1.1 Dinámica de entrenamiento Vs pérdida de validación 64
6. 1.2 Dinámica de entrenamiento Vs validación AUC 68
6. 2 Evaluación clínica comparativa 71
6. 2.1 Matrices de confusión 72
6. 2.2 Curvas ROC y Precision - Recall 75
6. 3 Interpretabilidad clínica mediante Grad-CAM 77
6. 3.1 Visualización del mejor experimento 80
6. 4 Comparación global entre arquitecturas 81
6.5 Discusión general de resultados 84
6. 5.1 Hallazgos principales 85
6. 5.2 Impacto clínico de FN y Recall 86
6. 5.3 Limitaciones experimentales 86
6. 5.4 Potencial clínico futuro 88
7. Conclusiones 90
8. Recomendaciones 92
9. Bibliografía. 93
10. Anexos 95

Lista de figuras

Pág.

Figura 1 32
Curva de realce cinético (TIC) Tipo 1: Patrón persistente 32
Figura 2 32
Curva de realce cinético (TIC) Tipo 2: patrones de meseta compuestos (Plateau) 32
Figura 3 33
Curva de realce cinético (TIC) Tipo 3: patrón de lavado. (Wash-out) 33
Figura 4 35
Categorías del Sistema BI-RADS, Definición y Probabilidad de Malignidad. 35
Figura 5 36
Arquitectura Típica de una Red Neuronal Convolucional 36
Figura 6 39
Arquitectura Vision Transformer 39
Figura 7 40
Arquitectura general de MobileViT y comparación con Vision Transformer tradicional. 40
Figura 8
Estrategia de transferencia de aprendizaje y ajuste fino progresivo aplicada sobre las arquitecturas ResNet50, EfficientNet-B3 y MobileViT-S. 56
Figura 9. 65
Curvas train loss vs validation loss 65
Figura 10 69
Curvas train AUC vs validation AUC 69
Figura 11 73
Consolidado de matrices de confusión de los cinco experimentos 73
Figura 12 76
Curvas ROC y Precision - Recall correspondientes al experimento cuatro. 76
Figura 13 80
Grad- CAM correspondiente al experimento cuatro. 80

Lista de tablas

Pág.

Tabla 1 45
Estratificación Clínica, Fases Temporales y Distribución de Subconjuntos del Dataset 45
Tabla 2 46
Análisis Estadístico Descriptivo del Volumen de Cortes (Slices) por Paciente 46
Tabla 3 66
Consolidado de pérdidas de entrenamiento y validación obtenidos durante los cinco experimentos 66
Tabla 4 70
Consolidado de AUC máximos por experimento 70
Tabla 5 77
Consolidado por experimento de curvas ROC y Precision - Recall (Average Precision) 77
Tabla 6 79
Consolidado Grad- CAM correspondiente a los 5 experimentos. 79
Tabla 7 84
Tabla comparativa global correspondiente al Experimento cuatro 84

Lista de anexos

.

Anexo 1 Repositorio en GitHub
Anexo 2 Base de datos BreastDM (ROI).

Resumen

En este trabajo se desarrolló y evaluó un flujo experimental basado en aprendizaje profundo para la clasificación binaria de lesiones mamarias benignas y malignas utilizando imágenes de resonancia magnética dinámica con contraste (DCE-MRI). Se empleó el dataset público BreastDM junto con representaciones multicanal derivadas de diferentes fases temporales del contraste. El desarrollo experimental incluyó estrategias de normalización, aumentación de datos y separación a nivel de paciente mediante Hold-Out. Adicionalmente, se implementaron las arquitecturas ResNet50, EfficientNet-B3 y MobileViT-S utilizando transfer learning y fine-tuning sobre la inicialización de ImageNet. La evaluación fue realizada mediante métricas clínicas y computacionales como sensibilidad, especificidad, F1-score, AUC-ROC, PR-AUC y Grad-CAM para interpretabilidad visual. Los resultados permitieron analizar diferencias en capacidad de generalización, estabilidad y focalización anatómica entre las arquitecturas implementadas.

Palabras clave: cáncer de mama, DCE-MRI, aprendizaje profundo, transferencia de aprendizaje, clasificación binaria

Abstract

This work developed and evaluated a deep learning pipeline for binary classification of benign and malignant breast lesions using Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI). The public BreastDM dataset was used together with multichannel representations derived from different temporal contrast phases. ResNet50, EfficientNet-B3, and MobileViT-S architectures were implemented using transfer learning and progressive fine-tuning over ImageNet pretrained weights. Evaluation was performed using sensitivity, specificity, F1-score, ROC-AUC, PR-AUC, and Grad-CAM for visual interpretability. The results allowed comparison of the generalization capability, stability, and anatomical focalization of the implemented architectures.

Keywords: breast cancer, DCE-MRI, deep learning, transfer learning, binary classification

Glosario

AUC-ROC: Área bajo la curva ROC utilizada para medir la capacidad general de discriminación entre clases de un modelo de clasificación.

Backbone convolucional: Bloque principal de extracción de características dentro de una arquitectura de aprendizaje profundo.

Biopsia: Procedimiento médico utilizado para extraer muestras de tejido con fines diagnósticos.

Calibración probabilística: Capacidad de un modelo para generar probabilidades consistentes con la frecuencia real observada de los eventos predichos.

Clasificación binaria: Tarea de clasificación en la que un modelo debe diferenciar entre dos categorías, como lesiones benignas y malignas.

Curva ROC: Representación gráfica de la relación entre sensibilidad y tasa de falsos positivos para diferentes umbrales de clasificación.

DCE-MRI: Resonancia magnética dinámica con contraste (Dynamic Contrast-Enhanced Magnetic Resonance Imaging), técnica de imagenología médica utilizada para analizar el comportamiento temporal del contraste en tejidos mamarios.

Deep Learning: Rama del aprendizaje automático basada en redes neuronales profundas capaces de aprender representaciones complejas a partir de grandes volúmenes de datos.

Dropout: Técnica de regularización utilizada para reducir el sobreajuste durante el entrenamiento de redes neuronales.

Especificidad: Métrica utilizada para medir la capacidad del modelo para identificar correctamente casos negativos o lesiones benignas.

Fine tuning: Proceso de ajuste fino realizado sobre capas profundas de una arquitectura preentrenada con el fin de especializarla en una tarea específica.

Focal Loss: Función de pérdida diseñada para priorizar muestras difíciles durante el entrenamiento de modelos de clasificación.

Falso negativo (FN): Caso clasificado incorrectamente como negativo por el modelo, aunque realmente corresponde a una lesión maligna.

Falso positivo (FP): Caso clasificado incorrectamente como positivo por el modelo, aunque realmente corresponde a una lesión benigna.

Grad-CAM: Técnica de interpretabilidad visual utilizada para identificar regiones relevantes de una imagen que influyen en la decisión tomada por un modelo de aprendizaje profundo.

Hold-Out: Estrategia de validación experimental basada en la separación fija de los datos en subconjuntos de entrenamiento, validación y prueba.

ImageNet: Base de datos de imágenes ampliamente utilizada para el preentrenamiento de arquitecturas de visión por computadora.

Interpretabilidad: Capacidad de analizar y comprender las decisiones tomadas por un modelo de inteligencia artificial.

Learning rate: Parámetro de entrenamiento que controla la magnitud de actualización de los pesos durante el proceso de optimización.

Lesión benigna: Alteración o masa mamaria que no presenta comportamiento cancerígeno ni capacidad de invasión hacia otros tejidos.

Lesión maligna: Alteración o masa mamaria asociada a comportamiento cancerígeno, caracterizada por crecimiento descontrolado y capacidad de invasión tisular.

Overfitting: Fenómeno en el cual un modelo aprende excesivamente los datos de entrenamiento, reduciendo su capacidad de generalización sobre datos no vistos.

Patch: Fragmento pequeño de imagen utilizado como unidad de entrada en arquitecturas Vision Transformer.

Pesos preentrenados: Parámetros aprendidos previamente por una arquitectura durante su entrenamiento sobre grandes conjuntos de datos como ImageNet.

PR-AUC: Área bajo la curva precisión-recall utilizada para evaluar el desempeño de modelos de clasificación en escenarios con posible desbalance de clases.

Región de interés (ROI): Área específica seleccionada dentro de una imagen médica para realizar análisis computacional o diagnóstico.

Resonancia magnética (RM): Técnica de imagenología médica basada en campos magnéticos y radiofrecuencia utilizada para obtener imágenes detalladas de estructuras internas del cuerpo.

Scheduler: Estrategia utilizada para modificar dinámicamente la tasa de aprendizaje durante el entrenamiento.

Sensibilidad: Métrica que evalúa la capacidad del modelo para identificar correctamente casos positivos, como lesiones malignas.

Transferencia de aprendizaje: Estrategia de entrenamiento en la cual un modelo previamente entrenado sobre grandes conjuntos de datos es reutilizado y adaptado para una nueva tarea específica.

Vision Transformer (ViT): Arquitectura basada en mecanismos de autoatención aplicada al análisis de imágenes mediante fragmentos denominados patches.

Wash-in: Fase temprana de captación rápida del contraste en estudios DCE-MRI, asociada al incremento inicial de intensidad en la lesión.

Wash-out: Disminución o pérdida de intensidad del contraste durante fases tardías del estudio DCE-MRI, frecuentemente asociada a lesiones malignas.

Contraste dinámico: Variación temporal de intensidad observada durante diferentes fases de captación del medio de contraste en estudios DCE-MRI.

Angiogénesis: Proceso de formación de nuevos vasos sanguíneos asociado frecuentemente al crecimiento tumoral y a lesiones malignas.

BI-RADS: Sistema estandarizado de clasificación radiológica utilizado para categorizar hallazgos en imágenes mamarias y estimar probabilidad de malignidad.

Carcinoma ductal invasivo: Tipo de cáncer de mama maligno caracterizado por invadir tejidos circundantes a partir de los conductos mamarios.

Captación de contraste: Incremento de intensidad observado en imágenes DCE-MRI debido a la absorción del medio de contraste por parte del tejido.

Fibroadenoma hipercelular: Lesión mamaria benigna que puede presentar alta vascularización y comportamientos de captación similares a lesiones malignas.

Gadolinio: Medio de contraste utilizado en resonancia magnética para analizar patrones de vascularización y comportamiento tisular.

Meseta (Plateau): Patrón cinético intermedio observado en curvas TIC, caracterizado por estabilización de intensidad tras la captación inicial del contraste.

Patrón persistente: Tipo de curva cinética asociada generalmente a lesiones benignas, caracterizada por aumento progresivo de intensidad en fases tardías.

Explosión de gradientes: Fenómeno que ocurre durante el entrenamiento de redes neuronales cuando los gradientes aumentan excesivamente de magnitud, generando inestabilidad en la actualización de los pesos y dificultando la convergencia del modelo.

Convergencia: Proceso mediante el cual un modelo de aprendizaje profundo alcanza estabilidad durante el entrenamiento, reduciendo progresivamente el error y mejorando su desempeño a lo largo de las épocas.

Introducción

El avance de la inteligencia artificial ha permitido el desarrollo de nuevas herramientas orientadas al análisis automatizado de imágenes médicas, especialmente en áreas relacionadas con visión por computadora y aprendizaje profundo. En los últimos años, estas técnicas han comenzado a utilizarse en tareas clínicas como clasificación, segmentación e interpretación de imágenes diagnósticas, mostrando resultados prometedores en diferentes campos de la medicina.

Dentro de la imagenología mamaria, la resonancia magnética dinámica con contraste (DCE-MRI) es una de las técnicas más utilizadas para la evaluación de lesiones mamarias debido a su alta sensibilidad diagnóstica y capacidad para analizar el comportamiento del contraste en diferentes fases temporales. Sin embargo, este tipo de estudios genera grandes volúmenes de información visual cuya interpretación puede resultar compleja, especialmente al analizar patrones de captación asociados a lesiones benignas y malignas.

En este contexto, el aprendizaje profundo representa una alternativa importante para apoyar procesos de análisis médico mediante modelos capaces de aprender automáticamente características relevantes a partir de imágenes. Arquitecturas como las redes neuronales convolucionales y los modelos híbridos basados en mecanismos de atención han demostrado un desempeño importante en tareas de clasificación visual, incluyendo aplicaciones relacionadas con imágenes médicas.

El presente trabajo desarrolla y evalúa un flujo experimental para la clasificación binaria de lesiones mamarias utilizando imágenes DCE-MRI y arquitecturas de aprendizaje profundo. Para ello, se empleó el dataset público BreastDM junto con representaciones multicanal derivadas de diferentes fases temporales del contraste. Durante el desarrollo experimental se implementaron estrategias de transferencia de aprendizaje, ajuste fino progresivo, calibración probabilística e interpretabilidad visual mediante Grad-CAM sobre las arquitecturas ResNet50, EfficientNet-B3 y MobileViT-S.

Finalmente, este trabajo busca analizar el comportamiento clínico y computacional de las arquitecturas implementadas, aportando una referencia experimental para futuras investigaciones relacionadas con inteligencia artificial aplicada a resonancia magnética mamaria y clasificación asistida de lesiones.

1. Planteamiento del problema

1.1 Descripción

El cáncer de mama es la neoplasia más frecuente en todo el mundo y una de las principales causas de mortalidad por cáncer en mujeres. Según estimaciones de (Global Cancer Observatory, s. f.), en 2022 se registraron aproximadamente 2.29 millones de casos nuevos y alrededor de 666 000 muertes asociadas a esta enfermedad.

En América Latina se reportaron cerca de 56 000 muertes relacionadas con el cáncer de mama durante 2022, mientras que en Colombia se registraron aproximadamente 4572 muertes y una incidencia significativa de nuevos casos.

La problemática también afecta de manera importante a mujeres jóvenes. En el rango entre 30 y 59 años, Latinoamérica presentó más de 111 000 casos nuevos y alrededor de 23 000 muertes asociadas al cáncer de mama. En Colombia, para este mismo grupo poblacional, se reportaron aproximadamente 9155 casos nuevos y 2016 muertes.

Estos datos evidencian la necesidad de fortalecer estrategias de apoyo al diagnóstico que sea realmente confiable, rápido y oportuno, que permita un acceso pronto al tratamiento y amplíe las probabilidades de diagnóstico a tiempo.

En este contexto, los modelos computacionales basados en inteligencia artificial y aprendizaje automático han surgido como herramientas prometedoras para apoyar el diagnóstico del cáncer de mama. Estas tecnologías permiten analizar grandes volúmenes de imágenes médicas y datos clínicos con altos niveles de precisión, facilitando la identificación temprana de hallazgos sospechosos que podrían pasar desapercibidos durante la evaluación convencional. De

esta manera, los sistemas inteligentes pueden complementar el trabajo de los profesionales de la salud, contribuyendo a mejorar la rapidez, consistencia y oportunidad de los procesos diagnósticos, especialmente en entornos con alta demanda asistencial o acceso limitado a especialistas.

1.2 Limitaciones de la mamografía

Aunque la mamografía es la herramienta de tamizaje estándar, tiene limitaciones significativas. Como se sabe, los falsos positivos son comunes, esto ocasiona ansiedad en la paciente y lleva a realizar pruebas de seguimiento y biopsias invasivas innecesarias, saturando de este modo el sistema de salud y afectando la calidad de vida de pacientes sin enfermedad maligna. Este problema frecuente se presenta regularmente en mamas densas, ya que la grasa y el cáncer, tienden a teñirse del mismo color y suelen confundirse a menudo y generar estos diagnósticos de forma errónea. Es importante recalcar qué debido a esto se vuelve más complicado para el radiólogo realizar un correcto diagnóstico y esto lleva a requerir otro proceso cómo lo es la biopsia por punción. En mamas densas, el cáncer se traspone con la mamografía, entre un color grisáceo claro y un blanco brillante, dificultando la detección de lesiones y generando un mal diagnóstico. Según datos de 2022 de la Sociedad Estadounidense de Cáncer (ACS), las mamografías de detección pueden pasar por alto uno de cada ocho casos de cáncer de mama, generalmente en mamas densas, lo que evidencia limitaciones diagnósticas importantes en mamas densas con un error de diagnóstico del 12%. Además, se cuenta también con cifras similares que proporciona el Instituto Nacional de Cáncer (NCI) señalando qué las mamografías de detección no detectan un 20%, haciendo qué uno de cada cinco diagnósticos de cáncer mamario no se detecten.

1.3 Complejidad de la resonancia magnética

La resonancia magnética dinámica con contraste es la modalidad con mayor sensibilidad, por lo que es crucial en pacientes de alto riesgo. Sin embargo, su complejidad se debe a la gran cantidad de información generada por los estudios DCE-MRI, los cuales combinan múltiples cortes anatómicos junto con diferentes fases temporales de captación del contraste. La interpretación de estas imágenes depende demasiado de la experiencia del radiólogo; esto genera una alta diferencia de criterios entre los expertos y una especificidad moderada que debe ser mejorada.

La mayor dificultad en la interpretación de resultados de la resonancia magnética, se encuentra en que la cinética del contraste de gadolinio no se mueve de forma uniforme. El radiólogo debe evaluar a la vez las tres fases de captación rápida (wash - in) y lavado (wash -out), las cuáles se confunden visualmente entre las lesiones malignas y los problemas benignos que reciben mucha sangre, como los fibroadenomas hipercelulares o focos de adenosis; además, el ojo humano no es capaz de medir que tan variado es el tumor por dentro en la distribución tridimensional del contraste. Esta limitación introduce una dificultad para evaluar las imágenes a simple vista y hace que diferentes radiólogos puedan dar diagnósticos muy distintos ante un mismo exámen. Al no tener una total certeza, los especialistas prefieren adoptar una postura muy preventiva y conservadora. Esto los lleva a clasificar la lesión con una sospecha más alta de la que realmente tiene, lo que termina ordenando una gran cantidad de biopsias innecesarias en pacientes que se convierten en falsos positivos.

1.4 Barreras de acceso y demoras diagnósticas en Colombia

El sistema de salud colombiano tiene problemas de oportunidad que afectan directamente el pronóstico de las pacientes con cáncer de mama. La falta de rapidez dentro de la ruta asistencial agrava la enfermedad y reduce las probabilidades de supervivencia.

Esta problemática se evidencia con datos que confirman la demora crítica en la atención:

1.4.1 Diagnóstico tardío

De acuerdo con López Plazas, 2025), expertos en salud pública señalan que en Colombia, siete de cada diez mujeres son diagnosticadas en etapas avanzadas, lo que reduce drásticamente las posibilidades de curación.

1.4.2 Demoras documentadas

Los expertos clínicos dicen que el retraso total al tratamiento no debe superar los tres meses o 90 días. Sin embargo, los estudios confirman que el sistema excede este margen:

En la investigación realizada por (Bonilla-Sepúlveda, 2022a), titulada "Retrasos en el diagnóstico y tratamiento del cáncer de mama en Medellín, Colombia", se halló que más del 80% de las pacientes sufrió un retraso total en el tratamiento, siendo esta demora mucho mayor para las pacientes en el régimen subsidiado.

Así mismo, un estudio de (Martínez-Pérez et al., 2020) realizado en cuatro centros oncológicos de Medellín halló que la mediana del tiempo transcurrido desde la identificación del problema por la paciente hasta la realización de la biopsia diagnóstica fue de 104,5 días, excediendo por más de un mes los límites clínicamente aceptables.

1.4.3 El aprendizaje profundo como base para futuras soluciones de ingeniería

La gran cantidad de información generada por la resonancia magnética de contraste dinámico, sumada a la necesidad de realizar diagnósticos rápidos, supera la capacidad de análisis del ojo humano para identificar de forma objetiva, patrones complejos de realce entre lesiones benignas y malignas. En este contexto, el aprendizaje profundo surge como una herramienta capaz de extraer automáticamente características claves a partir de las imágenes médicas. En este proyecto, se propone el uso de arquitecturas de aprendizaje profundo basadas en representación 2D multicanal, donde diferentes fases temporales de captación del contraste son integradas como canales de entrada para el modelo. Este enfoque permite capturar patrones dinámicos de realce de manera computacionalmente eficiente, facilitando el uso de técnicas de transferencia de aprendizaje y mejorando la estabilidad del entrenamiento en datasets médicos de tamaño limitado; para la validación experimental del proyecto se hizo uso de el benchmark público (Zhao et al., 2023a), especializado en clasificación de lesiones mamarias mediante DCE-MRI.

(Computer-aided detection of breast cancer on the Wisconsin dataset, 2022)

De esta manera, el sistema busca generar una clasificación automatizada, reproducible y consistente de las lesiones mamarias, funcionando como una herramienta de apoyo al especialista y contribuyendo a disminuir la variabilidad diagnóstica entre radiólogos. Además, el proyecto incorpora estrategias de validación patient-level orientadas a prevenir data leake entre conjuntos de entrenamiento y prueba, fortaleciendo la confiabilidad experimental de los resultados.

El clasificador, entonces, no busca reemplazar al experto, sino darle un "tercer ojo"

automatizado que le permita enfocar su atención de manera inmediata y precisa en los casos de mayor riesgo y no sólo eso, al automatizar la parte más compleja del análisis de imágenes, el radiólogo puede evaluar una mayor cantidad de exámenes en menor tiempo, esto agiliza significativamente el diagnóstico definitivo y podría apoyar estrategias de priorización clínica orientadas a disminuir tiempos de espera diagnóstica, y aumentar la eficiencia general de las clínicas, permitiendo atender a muchas más pacientes por día.

Por todo esto, la exploración de herramientas computacionales basadas en aprendizaje profundo, orientadas al apoyo del diagnóstico médico es de interés en la comunidad científica. En este contexto, la presente investigación analiza experimentalmente el potencial de diferentes arquitecturas de aprendizaje profundo, como base para futuros sistemas de asistencia al aplicados a clasificación binaria de lesiones mamarias mediante, Dicha optimización del flujo asistencial podría contribuir a mejorar los procesos de priorización diagnóstica y apoyar la toma de decisiones clínicas en escenarios de alta demanda asistencial entre la adquisición de la RM y la orden de punción, atacando directamente los más de 100 días de retraso documentados en el país mediante tutelas qué fallan a favor del paciente por el retardo continuo y dilatado de los procesos y transforma un desarrollo de ingeniería de software en una intervención que apoya directamente la supervivencia de la población femenina. (Bonilla-Sepúlveda, 2022)

Esta combinación de alta incidencia, impacto en población joven, las limitaciones de la mamografía, y la subjetividad de la resonancia magnética, confirman la necesidad de una solución de ingeniería pronta y con altos estándares de calidad que permitan evitar falsos positivos y falsos negativos y promuevan clasificación correcta de las lesiones mamarias.

1.5 Formulación

¿Cómo se puede diseñar, implementar y evaluar un flujo de trabajo de clasificación asistida por aprendizaje profundo para la clasificación de lesiones mamarias benignas y malignas en imágenes de resonancia magnética dinámica con contraste para contribuir al fortalecimiento de procesos de priorización diagnóstica en el contexto del sistema de salud colombiano?

2. Justificación

Este proyecto surge como una propuesta de investigación orientada a contribuir al análisis asistido de lesiones mamarias mediante técnicas de aprendizaje profundo aplicadas a imágenes de resonancia magnética mamaria. En el contexto Colombiano existe interés creciente en el desarrollo de herramientas computacionales de apoyo diagnóstico que permitan fortalecer los procesos de análisis en imagenología médica. Además, este trabajo representa una oportunidad académica para desarrollar y evaluar una metodología experimental basada en inteligencia artificial aplicada al análisis de imágenes médicas, bajo un enfoque de investigación rigurosa.

2.1. Justificación social y clínica

El cáncer de mama es un problema grave de salud pública a nivel mundial y nacional, debido a su alta incidencia y alta tasa de mortalidad. En Colombia, las dificultades asociadas al acceso oportuno al diagnóstico y tratamiento, continúan representando un desafío importante para el sistema de salud, En este contexto, surge la necesidad de fortalecer estrategias de apoyo al diagnóstico que contribuyan en la mejora de procesos de análisis, priorización clínica, así cómo la reducción de procedimiento médicos innecesarios.

2.1.1 Problema en los tiempos de diagnóstico:

Actualmente, existen demoras significativas en el diagnóstico y tratamiento del cáncer de mama que superan los límites clínicamente aceptables. La ausencia de herramientas automatizadas para la clasificación de lesiones en imágenes de resonancia magnética, contribuyen a esta brecha, impidiendo la priorización de casos de alta sospecha de malignidad. Este proyecto justifica su existencia al proponer el desarrollo y evolución de un modelo basado

en aprendizaje profundo orientado a la clasificación binaria que propone una herramienta experimental de apoyo diagnóstico orientada a fortalecer procesos de clasificación y priorización clínica.

El desarrollo posterior de un sistema de asistencia al diagnóstico (CAD) sería la evolución de éste proyecto, que podría contribuir al fortalecimiento de procesos de priorización diagnóstica magnética mamaria. Al ofrecer una clasificación automatizada y muy confiable de las lesiones mamarias en la resonancia magnética, el modelo permite priorizar los casos de alta sospecha de malignidad, apoyando la identificación temprana de casos con alta sospecha de malignidad que podría contribuir a fortalecer procesos de priorización diagnóstica.

2.1.2 Necesidad de un diagnóstico consistente:

La interpretación manual de las complejas imágenes de resonancia magnética mamaria, es evidentemente desafiante. Esta labor se caracteriza actualmente por una alta variabilidad interobservadora, (es decir, grandes diferencias de evaluaciones entre múltiples expertos). Esto significa que los radiólogos no siempre tienen la misma interpretación o las mismas mediciones cuándo analizan el mismo estudio, lo que conlleva a múltiples diagnósticos.

La justificación de esta investigación, radica en la necesidad de continuar explorando herramientas computacionales, basadas en el aprendizaje profundo que puedan servir en futuros desarrollos, como apoyo complementario al análisis del radiólogo, Esto resulta importante para reducir la diferencia de criterios que nace de la interpretación manual de las complejas imágenes DCE-MRI. Al ser más específicos, se puede disminuir la tasa de falsos positivos que se reportan en la mamografía y la RM, evitando así biopsias invasivas y costosas innecesarias, y

minimizando la ansiedad de las pacientes.

2.2. Relevancia tecnológica y científica

Este proyecto es importante en el contexto de la aplicación de técnicas de inteligencia artificial y aprendizaje profundo en imágenes médicas, en especial, en tareas de análisis y clasificación de lesiones mamarias mediante imágenes de resonancia magnética con contraste dinámico.

La investigación experimental aborda el desafío asociado a la complejidad de este tipo de imágenes médicas, permitiendo analizar el comportamiento de diferentes arquitecturas de aprendizaje profundo usadas a menudo en visión por computadora. En este sentido, la investigación aporta una evaluación experimental orientada al análisis de potencial de estas técnicas para la clasificación binaria de cáncer mamario y su futuro posible uso en la generación de un sistema de asistencia que se brinde como apoyo al radiólogo, para un diagnóstico clínico más rápido.

2.2.1 Representación multicanal de imágenes DCE-MRI:

El uso de técnicas de aprendizaje profundo se justifica como una estrategia capaz para el análisis de patrones complejos de realce en imágenes mediante representaciones 2D multicanal derivadas de distintas fases temporales del estudio. Este enfoque permite aprovechar técnicas de transferencia de aprendizaje sobre arquitecturas preentrenadas, utilizando imágenes correspondientes al área de interés (ROI), mejorando la estabilidad del entrenamiento y la capacidad de generalización en datasets médicos de tamaño limitado.

2.2.2 Evaluación experimental de arquitecturas de aprendizaje profundo

El principal aporte de este trabajo consiste en la evaluación comparativa de diferentes arquitecturas de aprendizaje profundo ampliamente utilizadas en visión por computadora e imágenes médicas, incluyendo redes neuronales convolucionales, y arquitecturas híbridas basadas en mecanismos de atención, adicional de aplicación de transferencia de aprendizaje en éstas, con el fin de optimizar las arquitecturas al problema actual. La investigación busca analizar el comportamiento de estas arquitecturas en tareas de clasificación binaria de lesiones mamarias, utilizando imágenes DCE-MRI.

Así mismo la investigación permite explorar la capacidad de los modelos para aprender patrones de representación asociados al comportamiento de las lesiones mamarias, particularmente en relación con las variaciones de captación de contraste de gadolinio presentes en diferentes fases temporales del estudio. Esto permite analizar el potencial de las técnicas de aprendizaje profundo para aprender representaciones asociadas a patrones de realce y lavado del contraste en imágenes orientadas a la clasificación binaria de lesiones mamarias.

2.2.3 Transferencia de tecnología:

La presente investigación resulta relevante como aproximación experimental al uso de arquitecturas de aprendizaje profundo aplicadas al análisis de imágenes médicas DCE-MRI. El desarrollo y evaluación de modelos de clasificación binaria de lesiones mamarias permite explorar el potencial de estas técnicas como base para futuros desarrollos orientados al apoyo computacional del análisis radiológico.

Asimismo, el enfoque metodológico implementado puede servir como referencia para

investigaciones posteriores relacionadas con inteligencia artificial aplicada a imagenología médica y apoyo al diagnóstico mamario..

2.3. Relevancia metodológica

A nivel metodológico, el proyecto es relevante porque permite evaluar el comportamiento de diferentes arquitecturas de aprendizaje profundo aplicadas a clasificación de lesiones mamarias mediante imágenes DCE-MRI.

Además, el trabajo aporta una aproximación experimental basada en separación a nivel de paciente y evaluación de modelos sobre conjuntos independientes de entrenamiento, validación y prueba, contribuyendo al análisis de técnicas de inteligencia artificial aplicadas a imágenes médicas.

3. Objetivos

3.1 Objetivo general

Implementar y evaluar arquitecturas de aprendizaje profundo para la clasificación binaria de lesiones mamarias benignas y malignas en imágenes de resonancia magnética dinámica con contraste dinámico (DCE-MRI) mediante estrategias de entrenamiento y evaluación experimental aplicadas al análisis de imágenes de imágenes médicas.

3.2 Objetivos específicos

1. Diseñar un proceso de preprocesamiento, normalización y representación multicanal para las imágenes de resonancia magnética dinámica con contraste, orientada a la preparación de datos para modelos de aprendizaje profundo aplicadas a clasificación binaria de lesiones mamarias.

2. Implementar y comparar diferentes arquitecturas de aprendizaje profundo, incluyendo modelos basados en redes neuronales convolucionales y arquitecturas híbridas basadas en mecanismos de atención para la clasificación binaria de lesiones mamarias en categoría benignas o malignos.

3. Evaluar el desempeño de las arquitecturas implementadas mediante métricas clínicas cómo Sensibilidad (Recall), Especificidad, F1-score, y el área bajo la curva (ROC), utilizando estrategias de validación experimental orientadas utilizando conjuntos independientes de entrenamiento, validación y prueba.

4. Marco teórico

El presente marco teórico establece los conceptos clave, los fundamentos tecnológicos y el soporte de la literatura necesarios para el diseño, la implementación y la evaluación del flujo de trabajo experimental propuesto. El proyecto integra tres pilares fundamentales: la Oncología mamaria, la Imagenología de resonancia magnética y las arquitecturas de aprendizaje profundo.

4.1. Antecedentes de la clasificación asistida por computadora en imagenología

4.1.1. Evolución de los sistemas CAD en mama

Los sistemas de Diagnóstico Asistido por Computadora nacieron como herramientas orientadas a apoyar la interpretación de imágenes médicas y reducir la variabilidad en la evaluación radiológica. Sus primeras aplicaciones en imagenología mamaria estuvieron enfocadas principalmente en mamografía digital, utilizando técnicas de procesamiento de imágenes y métodos clásicos de extracción manual de características para la detección de regiones sospechosas. (Computer-aided detection of breast cancer on the Wisconsin dataset, 2022)

Con el avance del aprendizaje automático y, posteriormente, del aprendizaje profundo, los sistemas CAD evolucionaron hacia modelos capaces de aprender representaciones directamente desde los datos de imágenes, disminuyendo la dependencia de técnicas manuales de ingeniería de características.

La incorporación de imágenes de resonancia magnética mamaria, particularmente estudios DCE-MRI, representó un nuevo desafío debido a la complejidad espacial y funcional de

este tipo de imágenes, así como a la presencia de múltiples fases temporales asociadas a la captación dinámica del contraste. En este contexto, las técnicas modernas de aprendizaje profundo han mostrado resultados prometedores para tareas de clasificación, detección y análisis asistido en imagenología mamaria. (Sheth & Giger, 2019)

4.1.2. El Paradigma del aprendizaje profundo en Radiología

Durante la última década, el aprendizaje profundo se ha consolidado como uno de los enfoques teóricos más utilizados en radiología e inteligencia artificial aplicada a imágenes médicas. A diferencia de los sistemas clásicos que dependían de técnicas de extracción manual de características, las redes neuronales profundas son capaces de aprender automáticamente las características más discriminatorias directamente de los datos brutos (píxeles o vóxeles). Esta capacidad ha permitido avances significativos en tareas de detección, segmentación y para esta investigación, la clasificación binaria de lesiones médicas. (Esteva et al., 2017)

4.2. Bases clínicas, radiológicas y principios del diagnóstico

4.2.1. Principios de la resonancia magnética dinámica con contraste

La resonancia magnética dinámica con contraste (DCE-MRI) es una técnica de imagenología médica utilizada para evaluar el comportamiento vascular de las lesiones mamarias mediante la administración de contraste basado en gadolinio. Los tumores malignos suelen presentar procesos de angiogénesis y neovascularización, lo que produce una mayor permeabilidad vascular y una captación más rápida del contraste en comparación con lesiones benignas. Estas variaciones en la dinámica de captación y lavado del contraste constituyen un

elemento importante para la evaluación diagnóstica en resonancia magnética mamaria

4.2.2. Naturaleza de los datos 3D y dinámica temporal en DCE-MRI

La resonancia magnética mamaria genera información con volumen tridimensional (3D) de las estructuras anatómicas. En estudios DCE-MRI, la adquisición de múltiples fases posteriores a la administración de contraste basado en gadolinio incorpora información temporal relacionada con la dinámica de captación del contraste.

Esta dinámica temporal permite analizar patrones de realce y lavado asociados al comportamiento vascular de las lesiones mamarias. en la práctica clínica, dichos patrones suelen representarse mediante curvas de intensidad-tiempo (Time-signal intensity curve - TIC) las cuales se clasifican comúnmente en tres tipos: persistente, meseta y lavado, siendo este último uno de los patrones más asociados a sospecha de malignidad. (Breast MRI Findings, s. f.)

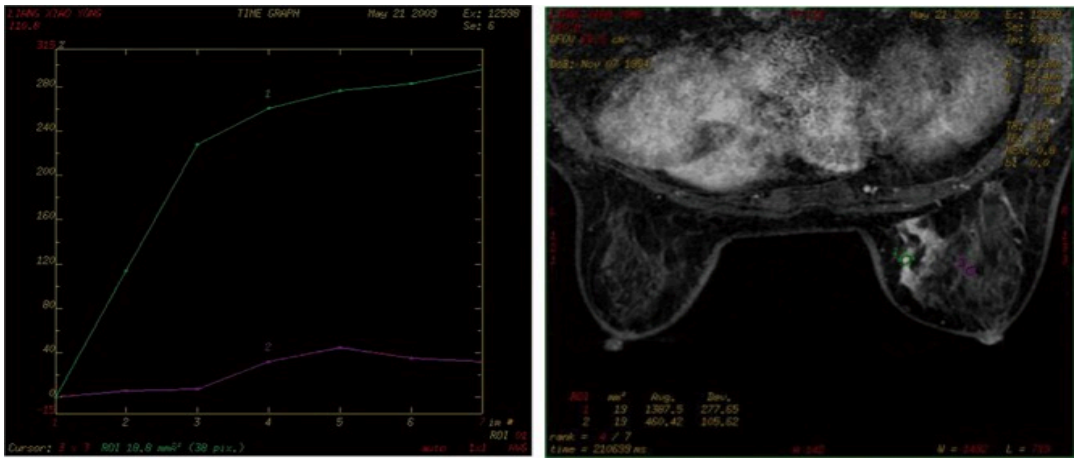
A continuación, se ilustran los principales patrones cinéticos reportados en la literatura especializada, los cuales representan características funcionales importantes en el análisis de imágenes DCE-MRI.

4.2.3 Curva persistente (Tipo 1: benigna) :

El Patrón Persistente es frecuentemente asociado a lesiones benignas y se caracteriza por un realce inicial rápido, seguido de una meseta o un ascenso continuo y lento en las fases tardías del estudio.

Figura 1

Curva de realce cinético (TIC) Tipo 1: Patrón persistente



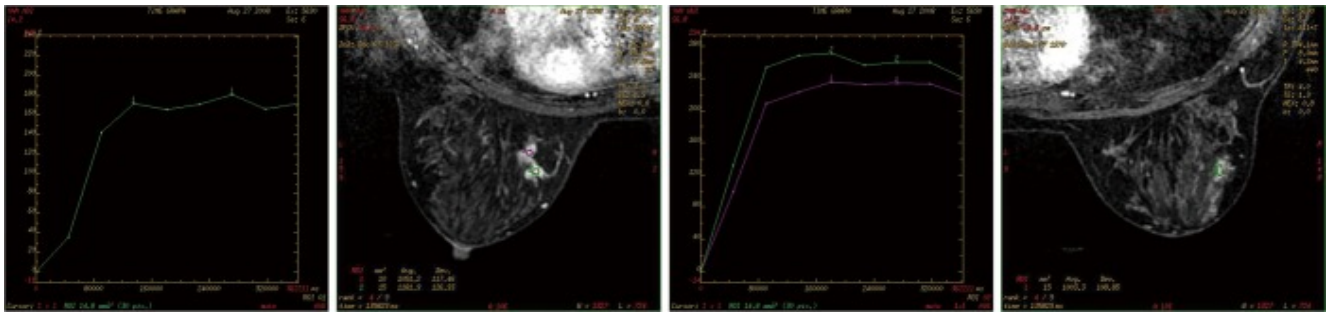
Nota. La curva muestra un realce persistente asociado a baja sospecha de malignidad. Tomado de Suo, J. et al. Dynamic Contrast-Enhanced Magnetic Resonance Imaging in the Differential Diagnosis of Breast Lesions. (2014).

4.2.4 Curva de meseta (tipo 2: sospechosa):

El Patrón de meseta representa una zona intermedia de sospecha. La curva alcanza su pico máximo rápidamente y luego se mantiene plana sin cambios significativos en las fases tardías. Este patrón no es concluyente por sí solo y a menudo requiere una evaluación más detallada.

Figura 2

Curva de realce cinético (TIC) Tipo 2: patrones de meseta compuestos (Plateau)



Tomado de Breast magnetic resonance imaging: kinetic curve assessment (Cheng & Li, 2013)

La figura 2 ilustra ejemplos del patrón cinético de meseta (Tipo 2), caracterizado por un realce inicial rápido seguido de una estabilización relativa en fases tardías del estudio. Este comportamiento suele considerarse un patrón intermedio o sospechoso en resonancia magnética mamaria, por lo que generalmente requiere una evaluación que complemente junto con las características morfológicas de la lesión.

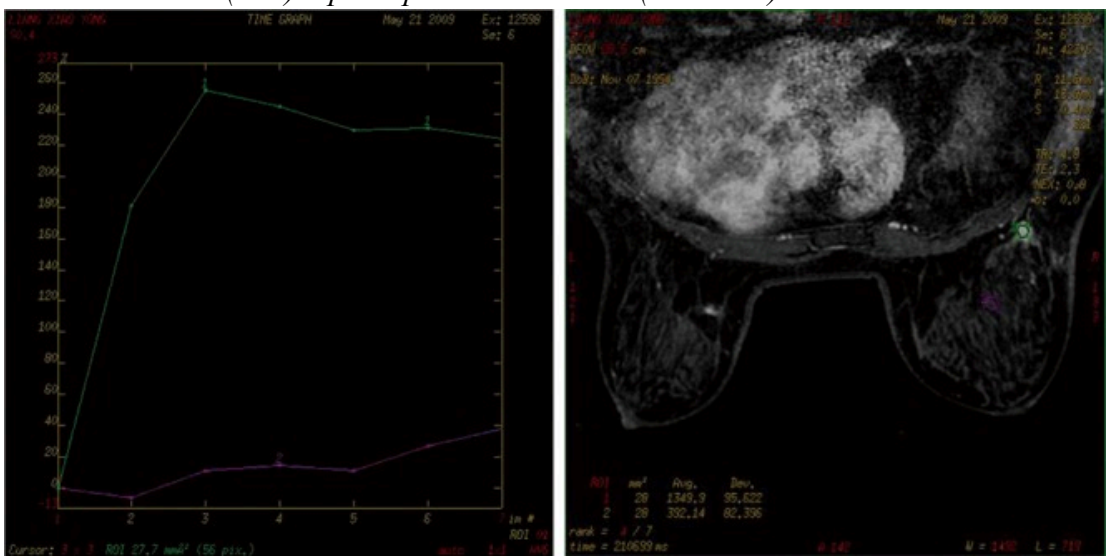
Las imágenes de resonancia magnética asociadas, muestran lesiones marcadas mediante regiones de interés (ROI), donde la combinación entre hallazgos morfológicos y comportamiento cinético contribuye al análisis diagnóstico.

4.2.5 Curva de lavado (Tipo 3):

El patrón de lavado (Tipo 3) es el más sugestivo de malignidad. Este comportamiento se debe a la alta permeabilidad de los vasos tumorales y a una rápida difusión del contraste fuera del tejido, resultando en un descenso significativo de la curva en las fases tardías del estudio.

Figura 3

Curva de realce cinético (TIC) Tipo 3: patrón de lavado. (Wash-out)



Tomado de Suo, J. et al. Dynamic Contrast-Enhanced Magnetic Resonance Imaging in the Differential Diagnosis of Breast Lesions. (2014)

La figura 3 compuesta por la gráfica cinética (panel izquierdo) y la imagen de resonancia magnética morfológica (panel derecho) ilustra el patrón más sugestivo de malignidad. La curva de realce (línea verde) presenta un pico máximo rápido e inmediatamente inicia un descenso brusco (lavado) en las fases tardías del examen. Este patrón cinético, sumado a las características morfológicas de la lesión visible en la resonancia magnética es altamente indicativo de Carcinoma ductal invasivo (Wash-out - Invasive breast carcinoma).

4.3. Sistema de reporte de datos de imagen mamaria

El sistema de reporte de datos de imagen mamaria de la American College of Radiology es el estándar universal para la categorización de hallazgos en imágenes mamarias, asignando una probabilidad de malignidad y un manejo clínico sugerido para cada lesión. La clasificación de la lesión por parte del modelo de aprendizaje profundo debe alinearse al objetivo clínico de asistir en la categorización BI-RADS 4 (Sospechoso) y BI-RADS 5 (Altamente Sugestivo de Malignidad), ya que estas categorías requieren una intervención de confirmación (biopsia).

La siguiente tabla resume las categorías principales del sistema, su definición y la probabilidad de malignidad asociada:

Figura 4

Categorías del Sistema BI-RADS, Definición y Probabilidad de Malignidad.

Categoría	Definición	Probabilidad de Cáncer
BI-RADS 0	Incompleto	N/A
BI-RADS 1	Negativo	Esencialmente 0%
BI-RADS 2	Benigno	Esencialmente 0%
BI-RADS 3	Probablemente Benigno	>0%, pero <2%
BI-RADS 4	Sospechoso	>2%, pero <95%
BI-RADS 5	Altamente sugestivo de malignidad	>95%
BI-RADS 6	Malignidad conocida por biopsia previa	N/A (Hay Cáncer diagnosticado)

Nota. Elaboración propia generada con Inteligencia artificial.

4.4. Fundamentos de ingeniería y enfoque teórico (aprendizaje profundo)

El proyecto se enmarca en la teoría del aprendizaje automático supervisado para tareas de clasificación, donde las imágenes médicas representan la entrada del modelo y el diagnóstico binario mamario.

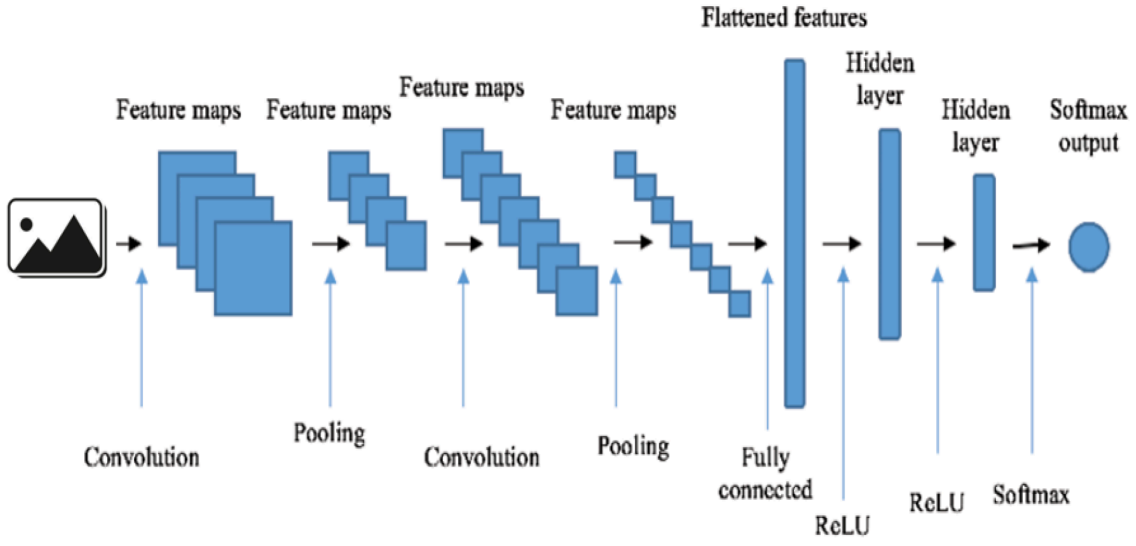
4.4.1. Teoría de las redes neuronales convolucionales (CNN)

Las redes neuronales convolucionales (CNN) constituyen uno de los enfoques más utilizados en visión por computadora y análisis de imágenes médicas. Su funcionamiento se basa en operaciones de convolución, donde diferentes filtros aprenden automáticamente patrones visuales de baja y alta complejidad, como bordes, texturas, formas y estructuras clave dentro de la imagen.

En este proyecto, las CNN son utilizadas sobre representaciones 2D multicanal derivadas de diferentes fases temporales de estudios DCE-MRI, permitiendo aprovechar información asociada a la dinámica de captación del contraste mediante estrategias de transferencia de aprendizaje sobre arquitecturas preentrenadas. Este enfoque facilita el aprendizaje de patrones significativos para la clasificación binaria de lesiones mamarias. (Litjens et al., 2017)

Figura 5

Arquitectura Típica de una Red Neuronal Convolucional



Tomada de (Convolutional Neural Networks and Vision Transformers for Fashion MNIST Classification: A Literature Review, s. f.)

La figura 5 ilustra el flujo de una CNN, que se divide en dos etapas principales:

1. **Extracción de Características:**

Esta fase está compuesta por la alternancia de capas de convolución (convolutional

layers) y de agrupación (pooling layers).

- **Convolución**: Aplica múltiples filtros (kernels o filters) que se desplazan sobre la imagen para generar mapas de características (feature maps). Esto permite al modelo detectar patrones jerárquicos esenciales, comenzando por elementos simples como bordes y texturas, hasta formas y estructuras más complejas.
- **Agrupación (Pooling):** Reduce la dimensionalidad espacial de los mapas de características generados. Su función es disminuir el costo computacional y mitigar el riesgo de sobreajuste (overfitting), manteniendo únicamente la información más relevante.

2. **Clasificación:**

Es la etapa final de la red donde se toma la información procesada para tomar una decisión.

- **Aplanado (Flattening)**: El mapa de características final, que es multidimensional, se transforma en un vector unidimensional .
- **Capas totalmente conectadas (Fully Connected)**: El vector aplanado se conecta a estas capas, las cuales procesan las características extraídas para aprender combinaciones de alto nivel.
- **Capa de salida (Output Layer)**: El flujo culmina en una función de activación Softmax, encargada de convertir las puntuaciones de la red en una distribución de probabilidad . Esto define la probabilidad final de pertenencia a cada clase predicha, en este caso, benigna o maligna.

4.4.2. Transformers visuales y mecanismos de atención

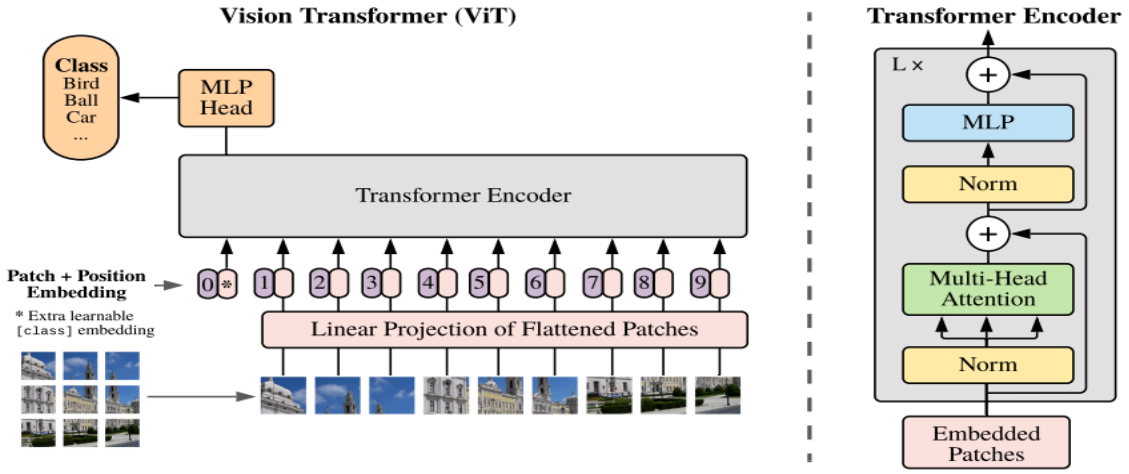
Las arquitecturas de Transformers visuales (ViT) se fundamentan en mecanismos de autoatención, propuestos originalmente por (Vaswani et al., 2017),los cuales permiten modelar relaciones entre diferentes regiones de una imagen mediante la asignación de distintos niveles de importancia a las características visuales durante el proceso de aprendizaje.

En visión por computadora, las redes neuronales convolucionales (CNN) han demostrado una gran capacidad para extraer características locales, mientras que los Vision Transformers (ViT) permiten modelar dependencias de largo alcance y capturar contexto global mediante mecanismos de atención. Por esta razón, han surgido arquitecturas híbridas que combinan ambas aproximaciones con el objetivo de aprovechar simultáneamente representaciones locales y globales de la imagen.

En el contexto de esta investigación, este enfoque resulta relevante debido a la implementación de MobileViT-S, una arquitectura híbrida entre CNN y Vision Transformer utilizada para la clasificación de lesiones mamarias en imágenes DCE-MRI. La combinación de características locales y contexto global puede contribuir al análisis de representaciones multicanal derivadas de diferentes fases temporales del contraste dinámico. (Kim et al., 2025)

4.4.3 Visión transformers (ViT) para la Imagenología

Figura 6
Arquitectura Vision Transformer



Tomada de (Dosovitskiy et al. 2021)

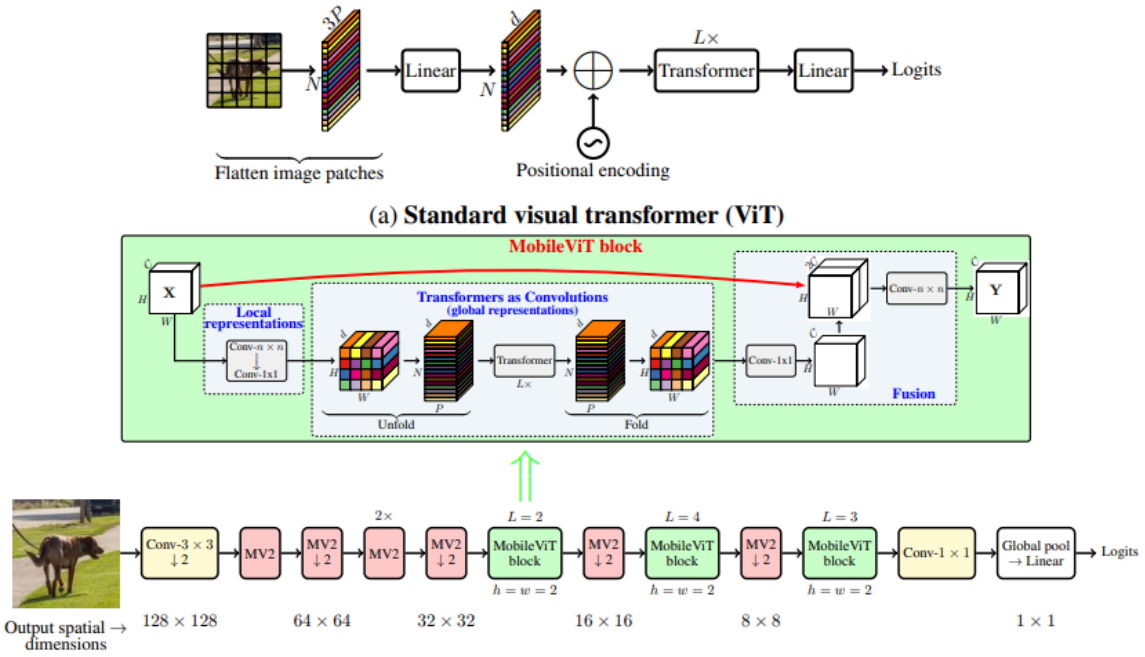
En aplicaciones de visión por computadora e imagenología médica, los modelos ViT han mostrado resultados prometedores en tareas de clasificación y análisis de imágenes complejas debido a su capacidad para modelar dependencias espaciales dentro de la imagen. (Khan et al., 2021)

Los Vision Transformers (ViT) corresponden a arquitecturas basadas en Transformer adaptadas al análisis de imágenes. A diferencia de las redes convolucionales tradicionales, estos modelos dividen la imagen en pequeños fragmentos denominados patches, los cuales son transformados en secuencias de embeddings procesadas mediante mecanismos de autoatención. Este enfoque permite modelar relaciones entre diferentes regiones de la imagen durante el proceso de aprendizaje (Dosovitskiy et al., 2021)

En este trabajo se emplea MobileViT-S, una arquitectura híbrida que combina capas convolucionales tradicionales para la extracción de características locales con módulos Transformer orientados al modelado de relaciones globales dentro de la imagen. Este enfoque busca integrar la capacidad de las CNN para detectar patrones espaciales de baja complejidad con el uso de mecanismos de autoatención para capturar dependencias de largo alcance. (Mehta & Rastegari, 2022)

Figura 7

Arquitectura general de MobileViT y comparación con Vision Transformer tradicional.



Nota. La figura ilustra la integración de bloques convolucionales y módulos Transformer dentro de la arquitectura MobileViT. En la parte superior se presenta el flujo general de un Vision Transformer tradicional, mientras que la sección central muestra el bloque híbrido MobileViT utilizado para combinar representaciones locales y contexto global durante el procesamiento de imágenes.
Tomado de (Mehta & Rastegari, 2022)

4.4.4 Atención global y representaciones multicanal

Al trabajar con representaciones multicanal derivadas de distintas fases de captación el contraste, los mecanismos de atención permiten analizar relaciones entre diferentes regiones de la imagen y patrones visuales asociados al realce de las lesiones mamarias. Este enfoque complementa la extracción de características locales realizada por las CNN tradicionales y representa una estrategia prometedora para tareas de clasificación en imagenología médica.

4.5 Métricas de evaluación de rendimiento

La validación experimental del proyecto se realizó mediante métricas ampliamente utilizadas en problemas de clasificación médica y aprendizaje automático, permitiendo evaluar la capacidad de los modelos para diferenciar lesiones benignas y malignas en estudios DCE-MRI.

4.5.1 Curva ROC y área bajo la curva (AUC)

La métrica ROC representa la relación entre la sensibilidad y la tasa de falsos positivos para diferentes umbrales de clasificación. El Área bajo la curva (AUC - ROC) mide la capacidad general del modelo para distinguir entre las categorías Benigno y Maligno.

Un valor de AUC cercano a 1.0 indica un alto poder discriminativo, mientras que un valor cercano a 0.5 representa un comportamiento similar al azar. (Fawcett, 2006)

4.5.2 Matriz de confusión y métricas clínicas

La matriz de confusión permite analizar el desempeño clínico de los modelos a partir de Verdaderos positivos (VP), Verdaderos Negativos (VN), Falsos Positivos (FP) y Falsos Negativos (FN). A partir de esta matriz, se derivan métricas fundamentales para la evaluación médica:

Sensibilidad (Recall): Evalúa la capacidad del modelo para detectar correctamente lesiones malignas.

Una sensibilidad es clínicamente importante para reducir la cantidad de falsos negativos.

Especificidad (Tasa de verdaderos negativos): Mide la capacidad del modelo para identificar correctamente lesiones benignas. Una alta especificidad contribuye a disminuir falsos positivos y procedimientos innecesarios.

Precisión: Evalúa la proporción de predicciones positivas que realmente corresponden a lesiones malignas.

F1-Score: Corresponde a la media armónica entre precisión y sensibilidad, siendo útil en problemas con posible desbalance de clases.

Adicionalmente, la evaluación experimental del proyecto se realizó bajo un esquema de validación preservando separación a nivel de paciente mediante un Método de exclusión (Hold-Out), evitando fuga de información entre pacientes y garantizando una evaluación más consistente del desempeño de los modelos. Finalmente, se emplearon métricas complementarias como el área bajo la curva precisión-recall (PR-AUC), Promedio Precisión (AP), valor predictivo positivo (PPV), valor predictivo negativo (NPV), tasa de falsos negativos (FNR) y Brier Score, las cuales permiten analizar el comportamiento probabilístico y clínico de los modelos. Estas métricas resultan útiles para evaluar la confiabilidad de las probabilidades generadas, la capacidad de calibración y el equilibrio entre detección correcta de lesiones malignas y reducción de errores diagnósticos

5. Metodología

La metodología aplicada se alineó con la naturaleza ingenieril y comparativa del proyecto, enfocándose en la evaluación experimental de modelos de aprendizaje profundo para clasificación binaria de lesiones mamarias en imágenes DCE-MRI.

5.1 Dataset y preparación de datos

En esta etapa se realizó la organización, preparación y estructuración de los datos utilizados durante el desarrollo experimental, incluyendo la selección de imágenes ROI, construcción de representaciones multicanal y separación de subconjuntos para entrenamiento, validación y prueba.

5.1.1 Descripción del dataset BreastDM

Para el desarrollo de esta investigación se utilizó el dataset BreastDM proporcionado por (Zhao et al., 2023), una base de datos médica compuesta por imágenes de resonancia magnética de mama con contraste dinámico (DCE-MRI) usadas para el análisis y clasificación de lesiones benignas y malignas.

La base de datos contiene estudios organizados por paciente e incluye diferentes secuencias temporales capturadas antes y después de la administración del medio de contraste de gadolinio. Estas secuencias permiten analizar cambios en la captación del contraste dentro del tejido mamario, característica que es importante en la detección y análisis de posibles lesiones tumorales.

El dataset final utilizado estuvo compuesto por 1722 slices correspondientes a 232

pacientes únicos, de los cuales hay 85 pacientes benignos y 147 pacientes malignos. Cada muestra fue organizada utilizando tripletas completas de imágenes DCE-MRI, conformadas por las secuencias precontraste (PRE), postcontraste temprano (EARLY) y postcontraste tardío (LATE).

5.1.2 Organización de imágenes ROI

Para el desarrollo de este experimento fue generado con ROI (Regiones de interés) previamente extraídas de la base de datos BreastDM. Los autores, además del artículo publicado, proporcionaron una versión del dataset ya tratada y lista para ser usada en tareas de clasificación binaria de cáncer mamario.

El uso de ROI ayudó a reducir información irrelevante presente dentro de las imágenes completas y permitió que las arquitecturas pudieran aprender características más representativas relacionadas con el tumor presentado en cada caso.

La clasificación usada dentro de este proyecto fue de tipo binaria, organizando las imágenes entre lesiones benignas y malignas de acuerdo con el etiquetado clínico brindado dentro del dataset.

Durante el proceso de preparación de los datos, se realizó la construcción y validación de archivos CSV correspondientes a entrenamiento, validación y prueba, garantizando una correcta organización a nivel de paciente*(Patient- Level)* de las imágenes usadas durante todo el flujo de trabajo experimental.

Cada muestra usada dentro del modelo fue construida mediante una representación multicanal compuesta por las tres fases temporales DCE-MRI: precontraste (PRE), postcontraste

temprano (EARLY) y postcontraste tardío (LATE). Esto permitió que cada slice anatómico fuera representado mediante una tripleta temporal completa, conservando la información dinámica del contraste utilizada posteriormente por las arquitecturas de aprendizaje profundo.

Aunque la cantidad de slices por paciente no fue uniforme dentro del dataset, todos los pacientes incluidos contaban con sus respectivas fases temporales completas, En este caso, el mínimo encontrado fue de 2 slices por paciente y el máximo de 31 slices, manteniendo siempre la representación multicanal PRE-EARLY-LATE para cada corte anatómico utilizado.

A partir de esto, se implementó una separación a nivel de paciente que evitara fugas de información entre entrenamiento, validación y prueba, garantizando ausencia de fuga de información entre subconjuntos

Para la división de los tres subconjuntos de datos, fue aplicado únicamente con los slices correspondientes a cada paciente dentro de su respectiva participación, usando una distribución Hold-Out 70/15/15 y distribuyéndolos según la cantidad de slices proporcionalmente en los tres subconjuntos.

A continuación, podrán visualizar éste esquema más a detalle:

Tabla 1

Estratificación Clínica, Fases Temporales y Distribución de Subconjuntos del Dataset

Subconjunto (Split)	Pacientes Unicos	Slices Totales	Casos Benignos	Casos Malignos	Fase PRE (VIBRANT)	Fase EARLY (VIBRANT+C2)	Fase LATE (VIBRANT+C7)
Entrenamiento	162 (69.83%)	1178 (68.41%)	59	103	1178	1178	1178
Validacion	35 (15.09%)	274 (15.91%)	13	22	274	274	274
Prueba	35 (15.09%)	270 (15.68%)	13	22	270	270	270
Dataset Global	**232 (100.0%)**	**1722 (100.0%)**	**85**	**147**	**1722**	**1722**	**1722**

Nota. Presenta la distribución clínica y estructural del dataset utilizado durante el desarrollo experimental y su separación bajo la estrategia hold-out.

Tabla 2

Análisis Estadístico Descriptivo del Volumen de Cortes (Slices) por Paciente

Subconjunto (Split)	Media (Mean)	Mediana (Median)	Desv. Estandar (Std Dev)	Minimo (Min)	Maximo (Max)
Entrenamiento (Train)	7.27	6.00	4.43	2	31
Validacion (Validation)	7.83	6.00	4.92	3	24
Prueba (Test)	7.71	6.00	5.18	2	22
Dataset Global (BreastD	**7.42**	**6.00**	**4.61**	**2**	**31**

Nota. Resume el comportamiento estadístico del número de slices por paciente dentro de cada subconjunto generado mediante la estrategia hold-out.

5.1.3 Representación multicanal DCE-MRI

Para la construcción de las muestras utilizadas durante el entrenamiento, cada imagen fue organizada mediante una representación multicanal compuesta por tres fases temporales provenientes de los estudios DCE-MRI.

En lugar de trabajar cada fase de manera independiente, las tres secuencias fueron integradas dentro de una única representación de entrada, permitiendo que los modelos analizaran simultáneamente el comportamiento temporal de captación de contraste dentro de la lesión mamaria.

La construcción multicanal fue realizada utilizando cada fase temporal como un canal independiente dentro de la imagen, de manera similar al funcionamiento de imágenes RGB. En este caso, cada canal representó un momento diferente del estudio DCE-MRI introduciendo así información temporal relacionada con la captación de contraste dentro del tejido mamario.

Las fases elegidas quedaron del siguiente modo:

para la fase PRE, la secuencia Vibrant, para la fase EARLY la secuencia temprana Vibrant + C2 y para la fase tardía LATE: Vibrant + C7. Esta combinación permitió construir representaciones multicanal capaces de conservar tanto información anatómica como cambios dinámicos asociados al comportamiento del contraste en las lesiones mamarias.

Además, esta representación facilitó la aplicación de técnicas de transferencia de aprendizaje sobre arquitecturas preentrenadas originalmente con imágenes de tres canales.

5.1.4 Normalización y preprocesamiento

Después de la organización de las imágenes ROI y de la construcción multicanal DCE-MRI, se aplicaron diferentes etapas de preprocesamiento con el objetivo de mejorar la estabilidad del entrenamiento y reducir variaciones presentes entre estudios médicos.

En un principio, las imágenes fueron normalizadas usando una estrategia Z-Score por estudio, permitiendo estandarizar la distribución de intensidades presentes dentro de cada secuencia. Esta decisión fue tomada debido a que las imágenes de resonancia magnética no manejan una escala absoluta de intensidad, ya que pueden existir variaciones dependiendo del escáner, protocolo de adquisición y paciente.

Debido a las variaciones de intensidad presentes entre estudios DCE-MRI y a los cambios producidos por la captación dinámica del contraste, se aplicó clipping basado en percentiles (p1-p99) con el fin de reducir la influencia de valores extremos y mejorar la estabilidad de la normalización.

Para evitar esto, el clipping basado en percentiles (p1-p99) permitió reducir la influencia

de estos picos hiperintensos, eliminando valores extremos de brillo y posibles artefactos de movimiento, comunes por la respiración de la paciente durante el estudio, evitando así que sesgan la distribución global de intensidades.

Al limitar los datos usando Z score, se previene que valores atípicos, como los picos hiperintensos generados por el contraste dinámico o artefactos propios del estudio, afecten negativamente el proceso de optimización, permitiendo mantener una mayor estabilidad numérica y una convergencia de gradientes más estable durante el entrenamiento de los modelos.

Posteriormente, cada arquitectura utilizó su resolución original de preentrenamiento. ResNet50 y MobileViT-S trabajaron con imágenes de 256×256 píxeles, mientras que EfficientNet-B3 utilizó imágenes de 300×300 píxeles, manteniendo así compatibilidad con la resolución originalmente utilizada durante el pretraining sobre ImageNet.

La decisión de conservar la resolución 300×300 para EfficientNet-B3 se tomó debido a que reducir la resolución de entrada podría afectar negativamente el proceso de transferencia de aprendizaje y el aprovechamiento de los pesos preentrenados de la arquitectura. Esto permite preservar compatibilidad con pesos transferidos, mantener estabilidad en la transferencia de aprendizaje y evita degradación de representaciones preentrenadas.

5.1.5 Estrategias de aumento de datos

Con el objetivo de mejorar la capacidad de generalización de los modelos y reducir el riesgo de sobreajuste, durante el entrenamiento se aplicaron diferentes técnicas de Data Augmentation sobre las imágenes ROI.

Las transformaciones utilizadas incluyeron rotaciones aleatorias controladas, traslaciones, escalamiento y volteo horizontal. Estas transformaciones fueron aplicadas dentro de rangos moderados, incluyendo rotaciones de 12°, traslaciones de 5 %, escalamiento entre 0.92 y 1.08, y volteo horizontal aleatorio de 0.5, con el fin de no alterar de manera agresiva la anatomía de las lesiones mamarias presentes en las imágenes.

Las estrategias permitieron aumentar la variabilidad visual del conjunto de entrenamiento sin modificar las características clínicas principales presentes dentro de las lesiones mamarias, favoreciendo así un aprendizaje más robusto por parte de las arquitecturas implementadas.

5.2 Tipo de proyecto

El presente proyecto se desarrolló como una investigación de tipo experimental y tecnológico, orientado al diseño, implementación y evaluación comparativa de arquitecturas de aprendizaje profundo para la clasificación binaria de lesiones mamarias a través de imágenes de resonancia magnética dinámica con contraste (DCE-MRI)

Desde el componente experimental, se realizó una evaluación completa usando diferentes arquitecturas aprendizaje profundo, sobre el mismo conjunto de datos, permitiendo analizar y comparar el comportamiento de clasificación de diferentes modelos basados en redes neuronales convolucionales y arquitecturas híbridas con mecanismos de atención.

Desde el componente tecnológico, el proyecto se enfocó en el desarrollo de un flujo de trabajo de clasificación asistida por IA orientado al análisis de regiones de interés (ROI) en estudios DCE-MRI de mama, introduciendo estrategias de transferencia de aprendizaje, ajuste fino de

hiperparámetros en cada modelo, ajuste fino de modelos (fine tuning), calibración probabilística e interpretación visual mediante GradCAM.

5.3 Método de investigación

El método de investigación usado combinó un enfoque experimental con análisis comparativo de diferentes arquitecturas de aprendizaje profundo aplicadas al diagnóstico asistido por imágenes médicas.

En un inicio, se realizó el preprocesamiento de las imágenes DCE-MRI obtenidas a través del dataset BreastDM, después del preprocesamiento, se implementó un flujo de trabajo de entrenamiento muy robusto, usando técnicas de transferencia de aprendizaje y fine tuning progresivo sobre distintas arquitecturas de aprendizaje profundo. Durante el desarrollo experimental también se realizaron ajustes de hiperparámetros y diferentes configuraciones de entrenamiento con el objetivo de mejorar el desempeño clínico de los modelos.

Finalmente, los modelos fueron evaluados usando métricas de clasificación clínica y herramientas de interpretabilidad visual como GradCAM, permitiendo analizar tanto el rendimiento cuantitativo como las regiones de atención usadas por cada arquitectura.

De este modo, se evaluó si los modelos realmente estaban aprendiendo correctamente de la imagen o si se estaban enfocando en regiones que no corresponden directamente con el tumor.

5.3.1 Procedimiento metodológico:

El procedimiento desarrollado para dar respuesta al problema y a los objetivos planteados se estructuró en las siguientes fases:

5.3.2 Fase de análisis y preprocesamiento:

Se realizó la organización y análisis de las imágenes DCE-MRI de la base de datos BreastDM (Zhao et al., 2023) Posteriormente, se aplicaron procesos de normalización, se seleccionó la base de datos tratada en regiones de interés (ROI) y generación de representaciones 2D multicanal derivadas de distintas fases temporales del contraste.

5.3.3 Fase de implementación y experimentación:

Se implementaron diferentes arquitecturas de aprendizaje profundo preentrenadas, incluyendo ResNet50, EfficientNet-B3 y MobileViT-S. Los modelos fueron entrenados utilizando estrategias de transferencia de aprendizaje y optimización experimental de hiper parámetros para la clasificación binaria de lesiones mamarias.

5.3.4 Fase de evaluación comparativa:

El desempeño de los modelos fue evaluado mediante métricas clínicas y de aprendizaje automático, incluyendo AUC-ROC, sensibilidad, especificidad, precisión, y F1-score. Para el desarrollo experimental el conjunto de datos fue organizado bajo una estrategia Hold-Out Validation, utilizando una distribución de 70% para entrenamiento, 15% para validación y 15% para prueba, Esta división permitió separar adecuadamente el proceso de entrenamiento, ajuste de hiper parámetros y evaluación final de los modelos.

Este enfoque permitió obtener una evaluación más robusta y consistente del rendimiento diagnóstico de las arquitecturas evaluadas.

Adicionalmente, se incorporaron estrategias de optimización de umbrales de decisión y análisis de interpretación mediante GradCAM para complementar la evaluación de los modelos.

5.3.5 Origen de los datos y tipología de fuentes

5.3.5.1 Fuentes primarias.

Imágenes médicas de resonancia magnética dinámica con contraste

Se obtiene de la base de datos pública BreastDM, desde el repositorio oficial de los autores (Zhao et al., 2023b) y publicada junto con el artículo : Esta base de datos proporciona estudios DCE-MRI de mama, e incluye entre varios elementos, las regiones de interés, asociadas a las regiones mamarias, proporcionadas en un zip y usadas para la clasificación y segmentación de imágenes

Códigos y datasets de implementación:

El software desarrollado, los scripts de preprocesamiento, y los conjuntos de datos de entrenamiento/prueba/validación se consideran fuentes primarias generadas durante el proyecto.

5.3.5.2 Fuentes secundarias:

Literatura científica:

Artículos en revistas indexadas con DOI, libros científicos y documentos de congresos sobre aprendizaje profundo en Radiología, CNNs, Transformers y DCE-MRI.

Bases de datos biomédicas:

Informes y publicaciones de la American College of Radiology (ACR), a Global Cancer Observatory (GLOBOCAN) y estudios epidemiológicos colombianos relacionados con el cáncer de mama y las demoras en la atención.

5.4 Arquitecturas implementadas

Para el desarrollo experimental del proyecto, se implementaron tres arquitecturas de aprendizaje profundo ampliamente utilizadas en tareas de visión por computadora e imágenes médicas: ResNet50, EfficientNet-B3 y MobileViT-S.

ResNet50 se utilizó como una arquitectura convolucional profunda basada en conexiones residuales, lo que permite mejorar el flujo de gradientes y facilitar el entrenamiento de redes de mayor profundidad. Esta arquitectura ha demostrado un comportamiento robusto en tareas de clasificación médica y fue una de las que presentó mayor estabilidad durante el desarrollo de cada uno de los experimentos realizados.

EfficientNet-B3 fue seleccionada debido a su estrategia de escalamiento compuesto, la cual permite balancear la profundidad, el ancho y la resolución de la imagen de entrada de manera eficiente. En este trabajo, la red operó utilizando imágenes de 300 x 300 píxeles, respetando su estructura original para preservar los pesos transferidos mediante transferencia de aprendizaje desde ImageNet.

Finalmente, se incorporó MobileViT-S, una arquitectura híbrida que combina operaciones convolucionales tradicionales con mecanismos de atención tipo Vision Transformer (ViT). La elección de este modelo se fundamenta en el sobresaliente desempeño que han reportado los ViT en la literatura reciente para el procesamiento de imágenes. Debido a que el proyecto trabaja con imágenes en 2D y no en 3D, se optó por esta variante híbrida, la cual da soporte a la propuesta de un enfoque multicanal que incorpora la dimensión temporal.

Todas las arquitecturas fueron adaptadas para tareas de clasificación binaria mediante la modificación de su capa final, permitiendo así diferenciar entre lesiones benignas y malignas.

5.4.1 Transfer learning y Fine-tuning

Debido al tamaño limitado del dataset médico usado durante el proyecto, el entrenamiento de las arquitecturas fue realizado usando estrategias de transferencia de aprendizaje. Cada modelo fue inicializado usando pesos preentrenados sobre ImageNet, permitiendo aprovechar características visuales previamente aprendidas durante el pre entrenamiento.

Inicialmente las capas convolucionales principales de cada arquitectura fueron congeladas, entrenando únicamente las capas finales de clasificación. Posteriormente se realizó un proceso de ajuste fino progresivo, descongelando parcialmente las capas profundas de los modelos para permitir un ajuste más especializado sobre las imágenes *DCE-MRI* utilizadas durante el experimento.

Esta estrategia permitió reducir el tiempo de entrenamiento, mejorar la estabilizador de convergencia y favorecer la capacidad de generalización de las arquitecturas dentro del problema de clasificación binaria de lesiones mamarias.

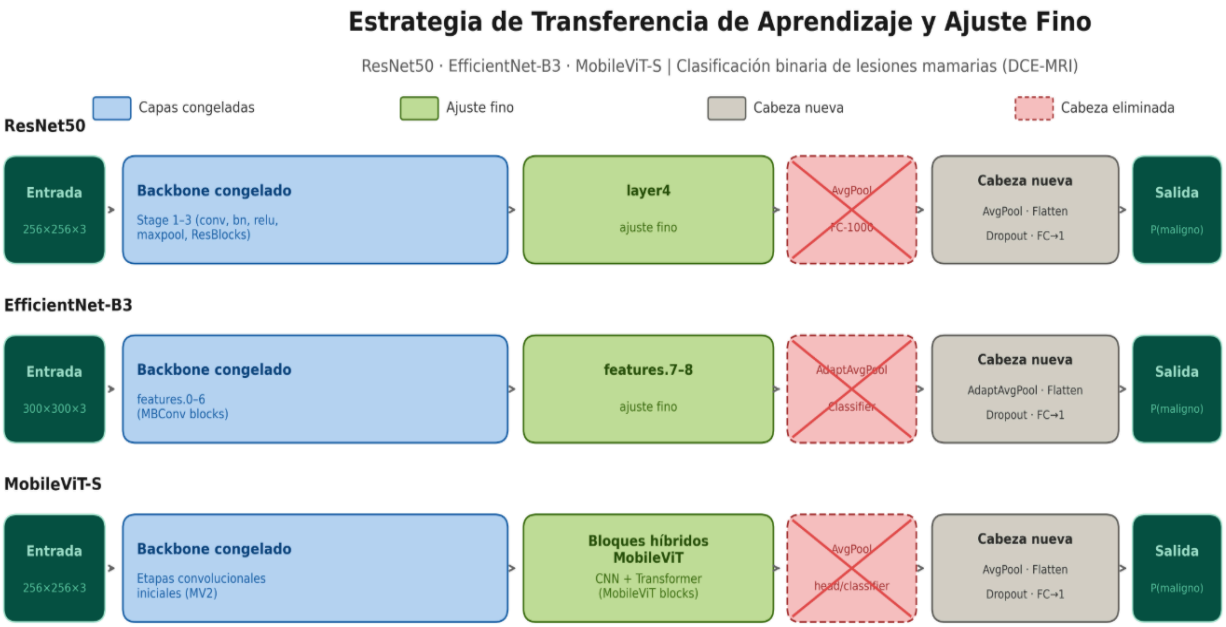
Durante la etapa inicial de transferencia de aprendizaje (transfer learning), únicamente se entrenaron las capas finales de clasificación de cada arquitectura, manteniendo congelado el bloque principal de extracción de características (backbone convolucional). En ResNet50 se entrenó inicialmente la capa completamente conectada (fully connected layer), en *EfficientNet-B3* la capa de clasificación (classifier) y en *MobileViT-S* las capas finales de clasificación (head/classifier), dependiendo de la estructura interna del modelo.

Posteriormente, se realizó un ajuste fino progresivo (fine tuning) sobre capas profundas específicas de cada arquitectura. En ResNet50 se habilitó el entrenamiento del bloque residual profundo (layer4); en EfficientNet-B3 se ajustaron parcialmente los bloques superiores correspondientes a las capas de extracción de características *(features.7 y features.8*); mientras que en MobileViT-S se realizó ajuste fino sobre los bloques profundos híbridos convolucionales-transformers. Esta estrategia permitió especializar progresivamente las representaciones aprendidas hacia el contexto clínico de imágenes DCE-MRI

Con el fin de ilustrar la estrategia de transferencia de aprendizaje y ajuste fino progresivo implementada durante el desarrollo experimental, la Figura 7 presenta las capas utilizadas durante las etapas de congelamiento y descongelamiento parcial aplicadas sobre las arquitecturas ResNet50, EfficientNet-B3 y MobileViT-S. Durante la primera fase se entrenaron únicamente las capas finales de clasificación utilizando pesos preentrenados sobre ImageNet y manteniendo congelado el backbone convolucional. Posteriormente, se realizó ajuste fino progresivo sobre capas profundas específicas de cada arquitectura con el fin de especializar las representaciones aprendidas hacia el contexto clínico de imágenes DCE-MRI.

Figura 8

Estrategia de transferencia de aprendizaje y ajuste fino progresivo aplicada sobre las arquitecturas ResNet50, EfficientNet-B3 y MobileViT-S.



Nota. Datos obtenidos a partir de configuraciones experimentales.

5.4.2 Configuración de entrenamiento

El entrenamiento de las arquitecturas fue realizado utilizando el optimizador AdamW junto con el scheduler CosineAnnealingLR para el ajuste dinámico de la tasa de aprendizaje durante las diferentes épocas de entrenamiento. En este contexto, los modelos computacionales buscaron optimizar progresivamente sus parámetros con el fin de mejorar la capacidad de clasificación sobre las imágenes médicas.

El optimizador AdamW fue seleccionado debido a su capacidad para desacoplar el decaimiento de pesos durante el proceso de optimización, contribuyendo a reducir el riesgo de sobreajuste y mejorar la estabilidad del entrenamiento lo que ayuda a que el modelo no

memorice y de este modo, no genere *overfitting*.

En cuanto al *scheduler* CosineAnnealingLR, permitió ajustar dinámicamente la tasa de aprendizaje a lo largo del entrenamiento, favoreciendo una convergencia más estable durante las diferentes épocas.

Cada modelo utilizó configuraciones específicas dependiendo de sus características arquitectónicas y resolución de entrada. ResNet50 y MobileViT-S trabajaron con imágenes 256 x 256 y batch size de 32, mientras que EfficienteNet - B3 usó imágenes 300 x 300 píxeles y batch size de 16, debido a su mayor consumo de memoria.

Durante el proceso, se implementó gradient clipping con un umbral máximo de 1.0. Esto funcionó como una medida de seguridad esencial para controlar el flujo de los gradientes, evitando que se "dispararan" (explosión de gradientes) y asegurando que el entrenamiento fuera numéricamente estable en cada época.

Para el entrenamiento de las capas finales de clasificación, se usó learning rates más altos durante la etapa inicial de entrenamiento, mientras que durante el ajuste fino progresivo, se empleó learning rates más bajos, permitiendo realizar ajustes más estables sobre los pesos preentrenados de las arquitecturas.

Adicionalmente, se implementó entrenamiento con precisión mixta (*mixed precision training*), permitiendo optimizar el uso de memoria GPU y reducir los tiempos de entrenamiento sin afectar el rendimiento de las arquitecturas.

5.4.3 Focal loss y optimización clínica

Como función de pérdida principal, se implementó Focal loss, usando un valor alpha de 0.6 y gamma de 2.0. Esta función fue seleccionada debido a su capacidad para penalizar de manera más fuerte las muestras difíciles durante el entrenamiento, favoreciendo el aprendizaje

sobre casos clínicamente complejos.

La adopción de pérdida focal estuvo orientada principalmente a mejorar el comportamiento clínico de las arquitecturas, especialmente en métricas relacionadas con sensibilidad (*Recall*) y reducción de falsos negativos (*FNR*), considerando que en el contexto médico resulta más crítico no detectar una lesión maligna que generar falsos positivos adicionales.

Durante el desarrollo experimental, las diferentes configuraciones de entrenamiento fueron evaluadas priorizando el desempeño clínico sobre el *Recall*, buscando reducir al máximo un diagnóstico erróneo sobre una persona con cáncer.

5.4.4 Calibración probabilística y ajuste de umbrales

Posteriormente al entrenamiento de las arquitecturas, se implementó estrategias de calibración probabilística usando calibración isotónica (*Isotonic Calibration)*. ajustando las probabilidades generadas por los modelos a partir del conjunto de validación y aplicándolas posteriormente sobre el conjunto de prueba.

Esta estrategia permitió obtener probabilidades más calibradas clínicamente, mejorando la interpretación de las predicciones generadas durante la clasificación binaria de lesiones mamarias.

Adicionalmente, durante la evaluación experimental se realizó optimización de umbrales de decisión, buscando configuraciones que permitieran favorecer métricas clínicas como la sensibilidad y reducción de falsos negativos.

La calibración y ajuste de umbrales, permitió complementar la evaluación cuantitativa

tradicional, orientando el comportamiento final de los modelos hacia escenarios más cercanos al contexto clínico real.

5.4.5 Entorno experimental y trazabilidad de experimentos

Con el objetivo de garantizar reproducibilidad computacional y mantener una adecuada trazabilidad experimental, durante el desarrollo del proyecto, se implementaron diferentes estrategias orientadas al control y seguimiento de los experimentos realizados.

Se empleó un entorno contenido en Docker con el fin de mantener un entorno de trabajo estable y reproducible, facilitando la instalación controlada de dependencias, librerías de aprendizaje profundo y herramientas usadas durante el entrenamiento. Además, esta estrategia permitió conservar la consistencia entre ejecuciones experimentales, evitando problemas relacionados con diferencia de versiones o cambios de configuración dentro del entorno de desarrollo.

Adicionalmente, durante el seguimiento experimental, se implementó herramientas de trazabilidad usando MLFlow, permitiendo registrar métricas, configuraciones, curvas de entrenamiento, hiperparámetros y resultados obtenidos en cada uno de los experimentos realizados.

Para mantener control sobre los cambios realizados durante el desarrollo del proyecto, se usó herramientas de control de versiones mediante Git, permitiendo organizar diferentes versiones del flujo de trabajo experimental desarrollados durante la investigación.

El uso de control de versiones facilita el seguimiento de modificaciones realizadas a lo largo de los distintos experimentos, ayudando a conservar respaldo del trabajo desarrollado y permitiendo recuperar configuraciones anteriores ante cambios en el rendimiento de los modelos.

5.5 Desarrollo experimental

El desarrollo experimental del proyecto se estructuró de manera iterativa, realizando múltiples configuraciones de entrenamiento y ajustes progresivos de hiperparámetros sobre cada una de las arquitecturas implementadas.

Durante este proceso, fueron evaluadas diferentes estrategias que iban desde cambiar la resolución de entrada de las imágenes y aplicar técnicas de regularización, hasta afinar las configuraciones de entrenamiento, la calibración probabilística y la optimización de los umbrales clínicos. Se probaron diversos modelos antes de tomar la decisión final de optar por las tres arquitecturas seleccionadas, ya que fueron las que demostraron el comportamiento más sólido frente al problema de clasificación binaria de las lesiones mamarias.

A lo largo de estas iteraciones, se fueron realizando ajustes metodológicos clave. La prioridad no era solo que el modelo memorizara los datos, sino mejorar la estabilidad del experimento y reducir sesgos de evaluación comunes en el área médica, como el sobreentrenamiento o el impacto del desbalance de clases. se buscó siempre configuraciones que ofrecieran un comportamiento clínicamente consistente y confiable.

Finalmente, las configuraciones con mejor desempeño fueron seleccionadas para la evaluación clínica definitiva, priorizando métricas relacionadas con sensibilidad diagnóstica, reducción de falsos negativos y capacidad de generalización frente a pacientes no observados durante el entrenamiento.

5.5.1 Métricas utilizadas

La evaluación experimental de las arquitecturas fue realizada usando métricas aprendizaje automático orientadas al análisis de clasificación binaria de lesiones mamarias.

Las principales métricas usadas fueron AUC - ROC, Sensibilidad (Recall), Especificidad, F1 - Score y la tasa de falsos negativos(FNR). Estas métricas permiten analizar tanto la capacidad discriminativa de los modelos como su comportamiento frente a lesiones.

Debido al contexto médico del proyecto, se priorizó especialmente el análisis de sensibilidad y FNR, considerando que los falsos negativos representan uno de los errores clínicamente más críticos dentro de tareas de clasificación de cáncer de mama.

5.5.2 Curvas ROC y Precision - Recall

Durante el desarrollo experimental se usó curvas ROC y curvas Precision- Recall para analizar el comportamiento de clasificación de las arquitecturas implementadas frente a diferentes umbrales de decisión.

Las curvas ROC permitieron evaluar la capacidad de discriminación entre lesiones benignas y malignas, mientras que las curvas Precision - Recall, permitieron analizar el análisis del comportamiento clínico de los modelos frente al equilibrio que hay entre sensibilidad y precisión diagnóstica, la precisión no se evalua sola, sino en conjunto.

Estas representaciones gráficas complementan la evaluación cuantitativa tradicional y permitieron comparar el comportamiento general de las arquitecturas elegidas y aplicadas para el desarrollo de este experimento.

5.5.3 Matrices de confusión

Las matrices de confusión fueron utilizadas para analizar el comportamiento clínico detallado de cada arquitectura implementada.

A partir de estas matrices, fue posible que identificará la cantidad de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos generados por cada modelo durante la evaluación sobre el conjunto de test.

Este análisis permitió complementar las métricas globales de desempeño y facilitó la interpretación clínica de los errores cometidos por las arquitecturas durante la clasificación binaria de lesiones mamarias.

5.5.4 Curvas de calibración

En cada experimento, también se hizo uso de las curvas de calibración probabilística con el objetivo de analizar la relación entre las probabilidades predichas por los modelos y el comportamiento real observado dentro del conjunto de prueba.

Este análisis es fundamental a nivel clínico, ya que permite auditar si los porcentajes de certeza entregados por el sistema son fiables; un modelo puede acertar en su diagnóstico binario (determinar si una lesión es maligna), pero exagerar a la vez, su nivel de confianza (por ejemplo, reportar un 99% de probabilidad cuándo probabilísticamente corresponde a un 65%), es por este motivo que se hizo uso de la calibración probabilística.

Para la correcta interpretación de estas métricas, el análisis visual se estructuró a partir de tres comportamientos respecto a la diagonal ideal de calibración perfecta:

- **Curva por encima de la diagonal**

  Indica que el modelo adopta una postura conservadora y subestima el riesgo, asignando

probabilidades de malignidad más bajas de lo que estadísticamente corresponde.

- **Curva por debajo de la diagonal**

Significa que el modelo sobreestima el riesgo y genera probabilidades de malignidad más altas de las reales. Clínicamente, este comportamiento probabilístico deficiente, tiende a producir un exceso de falsos positivos, lo que se traduce directamente en la ordenación de biopsias innecesarias para las pacientes.

- **Curva pegada a la diagonal**

Representa una calibración óptima donde las estimaciones del modelo no exageran ni atenúan la realidad del diagnóstico. En este escenario, el score probabilístico es altamente confiable y queda validado para su uso directo en la toma de decisiones clínicas.

5.5.5 Interpretabilidad mediante GradCAM

Finalmente, para complementar la evaluación cuantitativa, se implementó GradCAM como herramienta de interpretabilidad visual sobre las 3 arquitecturas entrenadas.
Esta técnica permitió visualizar las regiones de atención usadas por los modelos durante el proceso de clasificación, facilitando el análisis de las zonas de la imagen que tuvieron mayor influencia sobre las decisiones generadas por cada arquitectura.

El uso de GradCAM permitió evaluar si las arquitecturas dirigían su atención sobre las regiones asociadas a la lesión mamaria o si existían patrones de atención sobre zonas irrelevantes dentro de la imagen.

6. Resultados y discusión

6. 1. Resultados de entrenamiento y convergencia

Durante el desarrollo experimental fue posible observar diferencias importantes en el comportamiento de entrenamiento entre las arquitecturas implementadas. A lo largo de cinco experimentos realizados, cada modelo presentó patrones distintos de convergencia, estabilidad y capacidad de generalización, permitiendo analizar cómo las diferentes estrategias de optimización afectaron el aprendizaje sobre imágenes DCE-MRI multicanal.

En términos generales, ResNet50 fue la arquitectura con aprendizaje más rápido y agresivo, alcanzado valores muy altos de área bajo la curva (AUC) en pocas épocas y mostrando una fuerte capacidad para diferenciar lesiones benignas y malignas. Sin embargo, también fue el modelo que presentó señales más tempranas de sobreajuste, especialmente después de la época 12 en todos los experimentos, lo que sugiere una implementación de detección temprana (*Early stopping*) para que el modelo se estabilice y no memorice. Por otro lado, EfficientNet- B3 mostró un comportamiento mucho más estable y progresivo, manteniendo diferencias más pequeñas entre entrenamiento y validación, lo que evidenció una mejor capacidad de generalización global. Finalmente, MobileViT-S presentó un comportamiento intermedio, logrando convergencias relativamente estables, aunque con activaciones más variables y una tendencia moderada al sobreajuste en etapas tardías.

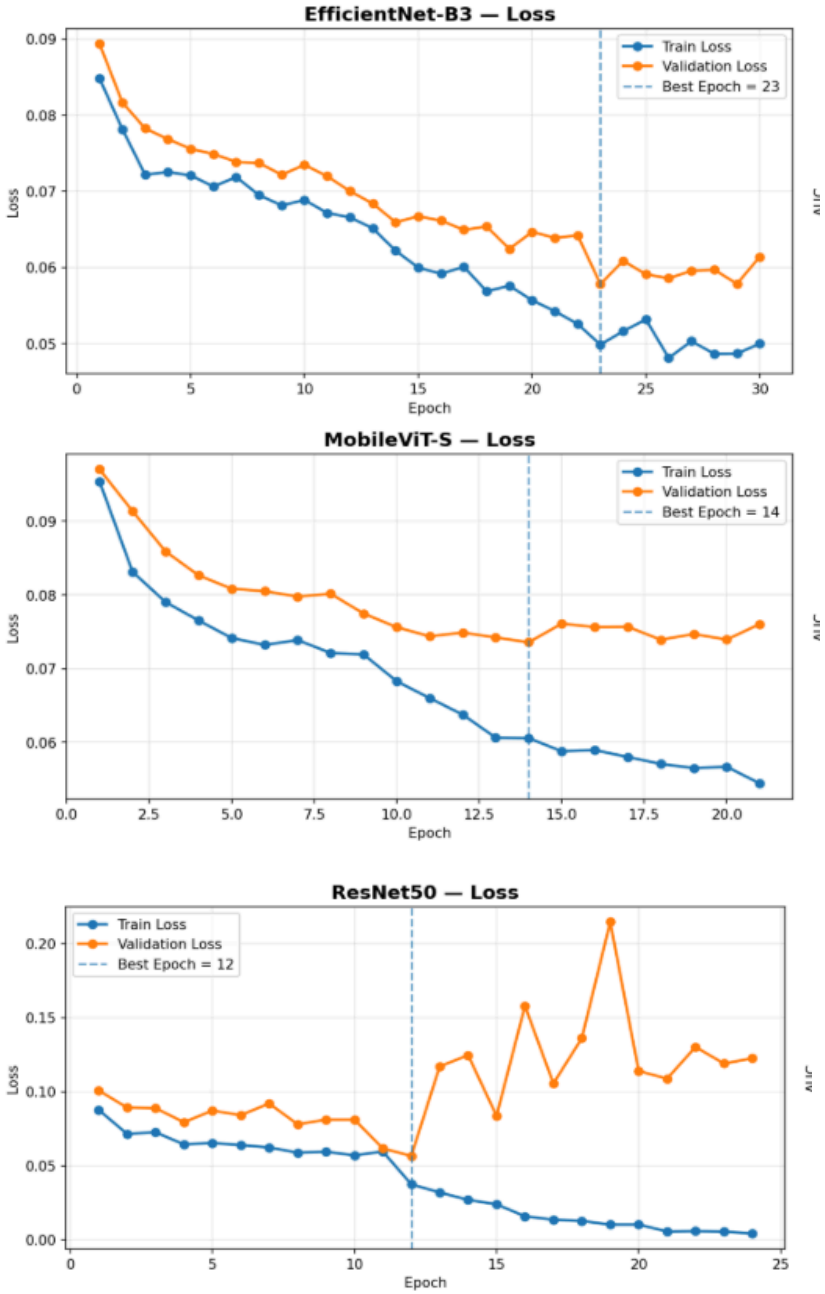
6. 1.1 Dinámica de entrenamiento Vs pérdida de validación

Las curvas de pérdida permitieron analizar el comportamiento de aprendizaje de las

arquitecturas a lo largo del entrenamiento, evaluando tanto la capacidad de ajuste sobre los datos de entrenamiento, como el comportamiento de generalización sobre el conjunto de validación.

Figura 9.

Curvas train loss vs validation loss



Nota. Esta imagen pertenece al cuarto experimento, el cuál corresponde al escenario con mejor rendimiento obtenido durante las pruebas.

Tabla 3

Consolidado de pérdidas de entrenamiento y validación obtenidos durante los cinco experimentos

Experimento	Modelo	Train Loss (Min)	Validation Loss (Min)	Diagnóstico
EXP 1	**ResNet50**	0.006	0.056	Excelente generalización inicial; overfitting después de época 12
	EfficientNet-B3	0.044	0.059	Generalización estable hasta época 23
	MobileViT-S	0.044	0.069	Entrenamiento estable; overfitting leve después de época 13
EXP 2	**ResNet50**	0.004	0.067	Generalización temprana óptima; overfitting después de época 12
	EfficientNet-B3	0.058	0.067	Generalización buena hasta época 23
	MobileViT-S	0.047	0.085	Overfitting moderado después de época 11
EXP 3	**ResNet50**	0.006	0.063	Aprendizaje rápido; overfitting después de época 12
	EfficientNet-B3	0.052	0.063	Generalización robusta hasta época 25
	MobileViT-S	0.044	0.079	Estabilidad moderada hasta época 14
EXP 4	**ResNet50**	0.004	0.056	Excelente generalización inicial; overfitting después de época 12
	EfficientNet-B3	0.050	0.058	Mejor generalización global hasta época 23
	MobileViT-S	0.054	0.074	Generalización aceptable hasta época 14
EXP 5	**ResNet50**	0.004	0.067	Generalización muy alta; overfitting después de época 12
	EfficientNet-B3	0.058	0.067	Generalización consistente hasta época 23
	MobileViT-S	0.047	0.085	Overfitting moderado después de época 11

De manera general, las arquitecturas lograron converger correctamente, mostrando disminuciones progresivas de la función de pérdida durante las primeras épocas. Sin embargo, el comportamiento posterior fue diferente entre modelos.

ResNet50 alcanzó los menores valores de train loss en prácticamente todos los experimentos, evidenciando una alta capacidad de aprendizaje y adaptación sobre las representaciones multicanal DCE-MRI. Sin embargo, en varios experimentos comenzó a observarse un incremento progresivo de pérdida de validación después de aproximadamente la época 12, indicando aparición de sobreajuste. A pesar de ello, antes de este punto, el modelo presentó uno de los mejores comportamientos de

generalización temprana de todo el estudio en todos los experimentos.

EfficientNet-B3 mostró el comportamiento más estable entre entrenamiento y validación. Sus curvas de pérdida descendieron de manera más gradual y consistente, manteniendo diferencias relativamente pequeñas entre la pérdida de entrenamiento y la pérdida de validación, incluso en épocas avanzadas. Este comportamiento permitió evidenciar una mejor estabilidad global y menor tendencia al sobreajuste, especialmente entre las épocas 23 y 31, dependiendo del experimento evaluado.

Por su parte, MobileViT-S presentó un comportamiento intermedio. Aunque logró convergencia estable durante las primeras épocas, posteriormente comenzó a evidenciar incrementos moderados en pérdida de validación, principalmente entre las épocas 11 y 14. Este comportamiento sugirió una capacidad de generalización aceptable, aunque menos estabilidad comparada con EfficientNet-B3, adicional, qué el modelo es muy grande para el tamaño de muestras.

En conjunto, el análisis de las curvas de pérdida y las métricas de validación permitió identificar diferencias importantes en la capacidad de generalización de las arquitecturas evaluadas. ResNet50 presentó el aprendizaje más rápido, aunque con señales más tempranas de sobreajuste. EfficientNet-B3 mostró el comportamiento más estable y consistente durante el entrenamiento, evidenciando la mejor capacidad de generalización global. Por su parte, MobileViT-S alcanzó un comportamiento intermedio, manteniendo resultados competitivos aunque con menor estabilidad frente a EfficientNet-B3.

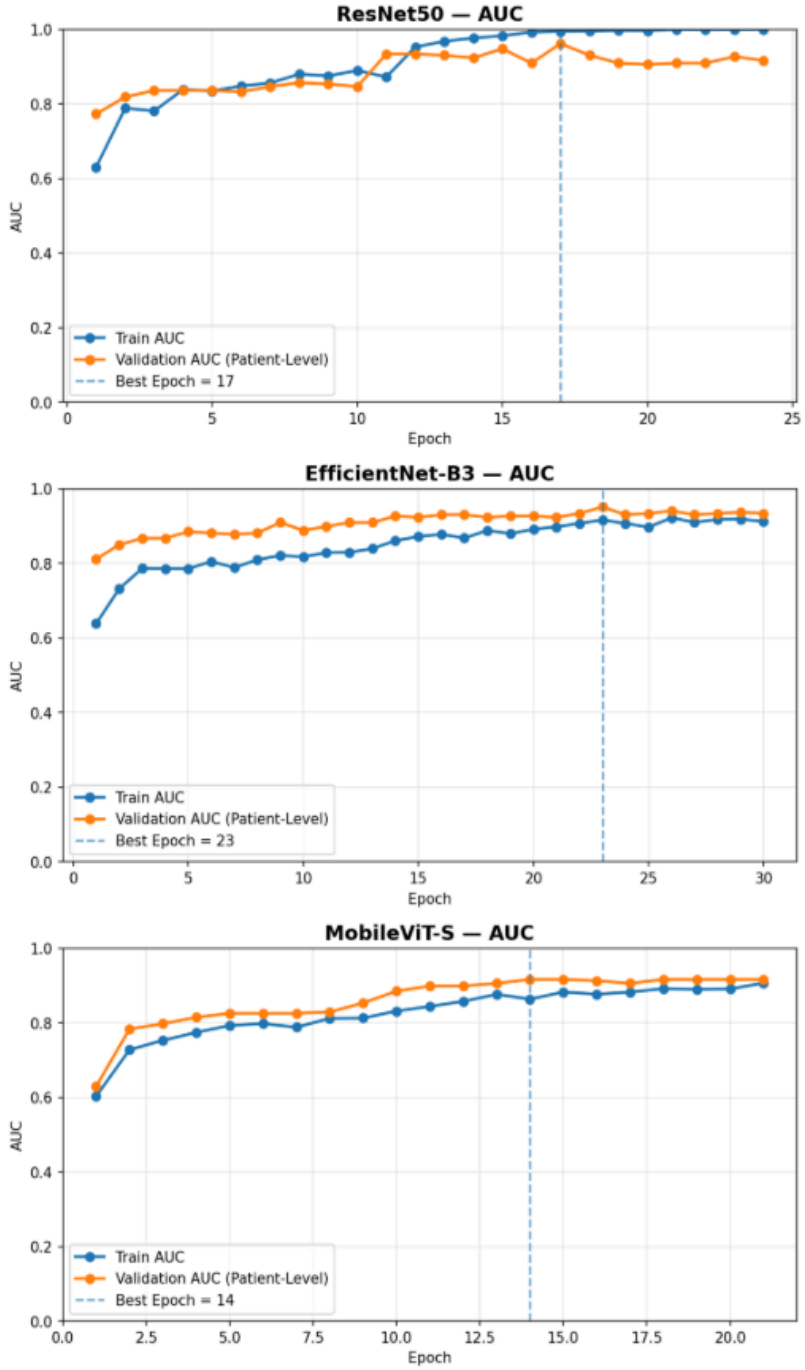
A nivel general, los resultados demostraron que la combinación entre representación multicanal DCE-MRI, normalización Z-Score, estrategias de aumentación de datos y técnicas de transferencia de aprendizaje permitió entrenar modelos robustos y clínicamente competitivos sobre un dataset médico relativamente pequeño, manteniendo estabilidad experimental y adecuada capacidad discriminativa durante el proceso de entrenamiento.

6. 1.2 Dinámica de entrenamiento Vs validación AUC

Las curvas AUC permitieron analizar la capacidad de discriminación de los modelos durante el proceso de entrenamiento, evaluando qué tan bien lograban separar las clases benignas y malignas a medida que avanzaban las épocas.

Figura 10

Curvas train AUC vs validation AUC



Nota. Esta imagen pertenece al cuarto experimento, el cuál corresponde al escenario con mejor rendimiento obtenido durante las pruebas.

Tabla 4

Consolidado de AUC máximos por experimento

Experimento	Modelo	Train AUC (Max)	Validation AUC (Max)
EXP 1	ResNet50	1.00	0.95
	EfficientNet-B3	0.93	0.94
	MobileViT-S	0.93	0.92
EXP 2	ResNet50	1.00	0.95
	EfficientNet-B3	0.92	0.95
	MobileViT-S	0.95	0.91
EXP 3	ResNet50	1.00	0.94
	EfficientNet-B3	0.92	0.95
	MobileViT-S	0.95	0.91
EXP 4	ResNet50	1.00	0.96
	EfficientNet-B3	0.92	0.95
	MobileViT-S	0.91	0.91
EXP 5	ResNet50	1.00	0.96
	EfficientNet-B3	0.92	0.95
	MobileViT-S	0.95	0.91

Los resultados de la tabla 4, muestran que Resnet50 alcanzó los valores más altos de entrenamiento AUC en todos los experimentos, llegando a valores cercanos a 1.00 en los cinco experimentos. Esto evidencia una alta capacidad de aprendizaje sobre los patrones dinámicos presentes en las imágenes DCE-MRI. Sin embargo, aunque el modelo logró excelentes resultados discriminativos, las diferencias crecientes entre Train AUC y Validation AUC en algunas etapas del entrenamiento, confirmaron nuevamente la aparición de sobreajuste en épocas mayores a la época doce, por lo que se sugiere hay que detener el entrenamiento allí con esta base de datos tan pequeña.

EfficientNet-B3 por otro lado, presentó curvas *AUC* más suaves y progresivas, manteniendo un comportamiento mucho más consistente entre entrenamiento y validación, aunque sus valores máximos de entrenamiento AUC fueron ligeramente inferiores a ResNet50, el modelo logró mantener estabilidad discriminativa sostenida durante un mayor número de épocas, evidenciando una mejor

capacidad de generalización global.

Por otra parte, MobileViT-S Mostró resultados competitivos, aunque con mayor variabilidad entre experimentos, sus curvas *AUC* reflejaron una adecuada capacidad discriminativa durante las primeras etapas de entrenamiento, pero posteriormente se observó fluctuaciones más marcadas entre entrenamiento y validación, indicando mejor estabilidad frente a cambios del conjunto de datos.

En términos generales, las curvas AUC permitieron confirmar que las representaciones multicanal construidas a partir de las fases pre-contraste, realce temprano y realce tardío, fueron capaces de aportar información discriminativa relevante para la clasificación binaria de lesiones mamarias, permitiendo que las arquitecturas aprendieran patrones asociados al comportamiento del contraste.

6. 2 Evaluación clínica comparativa

Después de analizar el comportamiento de entrenamiento y convergencia de las arquitecturas implementadas, se realizó una evaluación clínica comparativa usando métricas orientadas al análisis diagnóstico de lesiones mamarias. Esta etapa permitió evaluar la capacidad real de los modelos para diferenciar correctamente entre lesiones benignas y malignas del dataset, considerando no solo el desempeño global, sino también el impacto clínico asociado a errores de clasificación como falsos positivos y falsos negativos.

Los resultados obtenidos mostraron diferencias importantes entre arquitecturas particularmente en sensibilidad, estabilidad de clasificación y comportamiento probabilístico, Resnet presentó el mejor desempeño clínico global, alcanzando los menores niveles de falsos negativos y mostrando alta capacidad discriminativa, EfficientNet-B3 mostró el comportamiento más estable y balanceado entre sensibilidad y especificidad mientras que MobileViT-S, mantuvo resultados aceptables, aunque con

menor precisión anatómica y mayor dependencia de contexto.

6. 2.1 Matrices de confusión

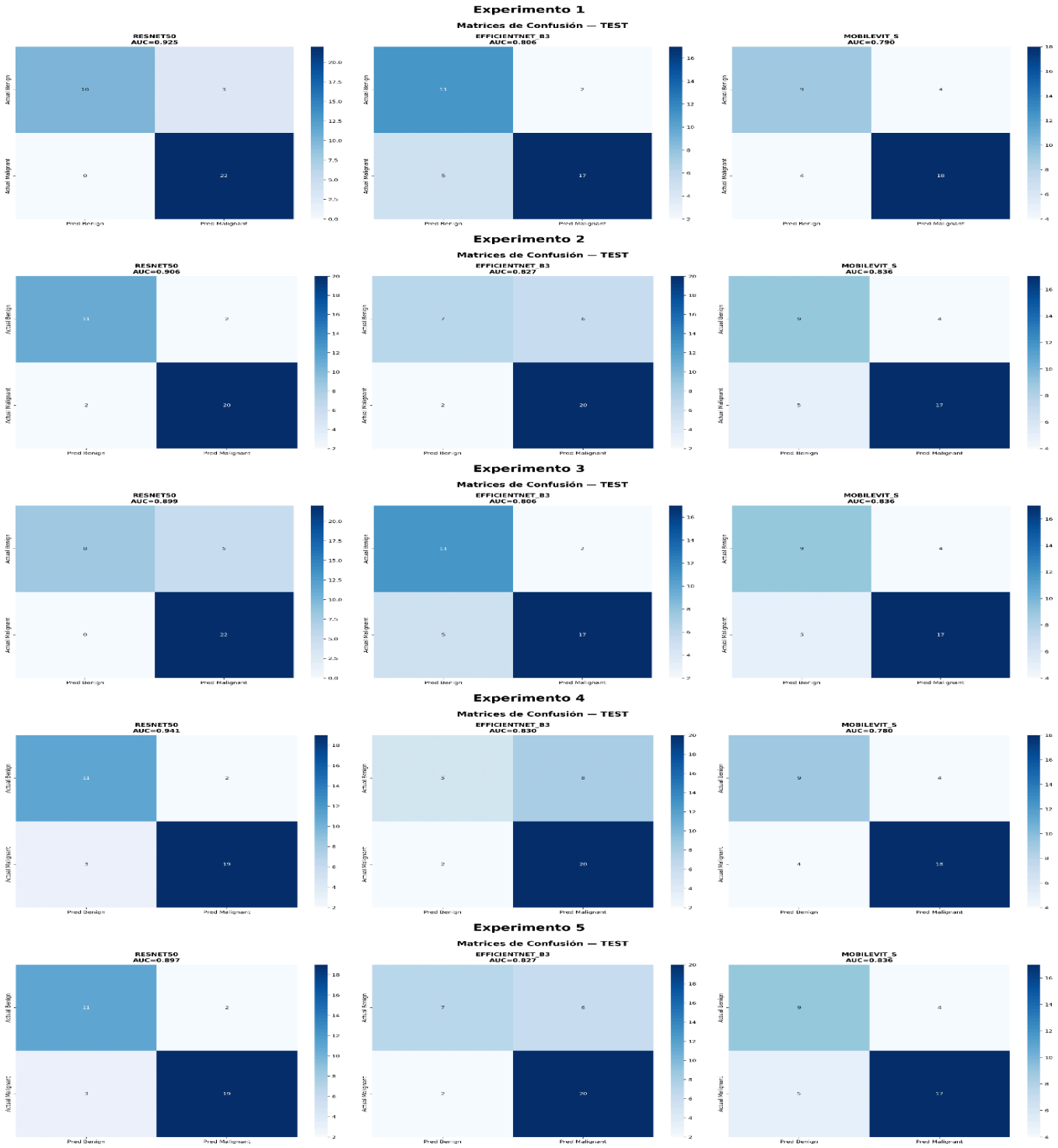
Las matrices de confusión permitieron analizar a detalle el comportamiento clínico de clasificación de las arquitecturas implementadas, evaluando la cantidad de verdaderos positivos, verdaderos negativos, falsos positivos y falso negativos obtenidos en cada experimento.

En el experimento 4, ResNet50 presentó el mejor desempeño general en el conjunto de pruebas, obteniendo 19 verdaderos positivos y 11 verdaderos negativos, con únicamente 3 falsos negativos y 2 falsos positivos, alcanzando un AUC de 0.941. EfficientNet-B3 logró detectar correctamente 20 lesiones malignas, aunque presentó una mayor cantidad de falsos positivos (8), obteniendo un AUC de 0.830. Por su parte, MobileViT-S alcanzó 18 verdaderos positivos y 9 verdaderos negativos, con 4 falsos negativos y 4 falsos positivos, registrando un AUC de 0.780.

En la figura 11, veremos más a detalle las matrices de confusión para cada experimento realizado

Figura 11

Consolidado de matrices de confusión de los cinco experimentos



En términos generales, los modelos lograron una adecuada capacidad de clasificación sobre lesiones benignas y malignas, Sin embargo, las diferencias más importantes entre arquitecturas

aparecieron principalmente en la cantidad de falsos negativos generados durante las pruebas.

ResNet50 presentó el comportamiento clínico más sólido de todo el estudio, alcanzando configuraciones donde logró reducir los falsos negativos a cero en algunos experimentos. Este resultado es importante en el contexto médico debido a que un falso negativo implica clasificar erróneamente una lesión maligna como benigna, aumentando el riesgo de retrasar el diagnóstico y tratamiento a tiempo del cáncer de mama.

Además de su alta sensibilidad, ResNet50 mostró matrices de confusión más consistentes y balanceadas, manteniendo una adecuada capacidad de detección, incluso bajo diferentes configuraciones experimentales.

EfficienteNetB3 presentó un comportamiento más estable entre experimentos, manteniendo balances relativamente consistentes entre verdaderos positivos y verdaderos negativos. Aunque en algunos casos mostró una tendencia ligeramente superior a generar errores de clasificación en comparación con las otras arquitecturas, esto debido a que es un modelo muy grande con muy pocos datos de entrenamiento, generando escenarios clínicamente ambiguos.

Por su parte, MobileViT-S mostró resultados competitivos en detección de lesiones malignas, aunque presentó una mayor variabilidad en la distribución de errores entre experimentos. Aun así, el modelo logró evidenciar la capacidad de las arquitecturas híbridas basadas en mecanismos de atención para aprender patrones relevantes asociados al comportamiento dinámico del contraste en imágenes DCE-MRI.

En general, las matrices de confusión, permitieron confirmar que las arquitecturas implementadas fueron capaces de aprender patrones clave asociados al comportamiento dinámico del contraste en imágenes DCE-MRI, logrando desempeños clínicamente competitivos para tareas de

clasificación binaria de lesiones mamarias.

6. 2.2 Curvas ROC y Precision - Recall

Las curvas ROC y Precision Recall permitieron analizar la capacidad de discriminación de las arquitecturas implementadas bajo diferentes umbrales de clasificación, evaluando la relación entre sensibilidad, especificidad y precisión durante el proceso de predicción.

ResNet50 alcanzó las curvas ROC más altas y consistentes, evidenciando excelente capacidad para separar las lesiones benignas y malignas. El mejor experimento sin duda, fue el experimento cuatro, alcanzando una ROC Curve de 0.941 y un Recall - Precision (AP) de 0.940.

EfficienteNet-B3 por su parte, mostró resultados muy competitivos con mejor estabilidad global, siendo el experimento cuatro, también el mejor, con una ROC Curve de 0.830 y una Precisión-Recall de 0.858.

Finalmente, MobileViT-S presentó mayor variabilidad entre experimentos, siendo el experimento 5, el mejor de ellos, con un resultado para ROC Curve de 0.836 y una Precision - Recall de 0.849.

En las curvas Precision - Recall fue posible observar que ResNet50 mantuvo el mejor equilibrio entre precisión y sensibilidad, aspecto especialmente importante en escenarios médicos con riesgo asociado a falsos negativos.

Estas curvas podrán ser vistas a detalle en la figura 12.

Figura 12

Curvas ROC y Precision - Recall correspondientes al experimento cuatro.

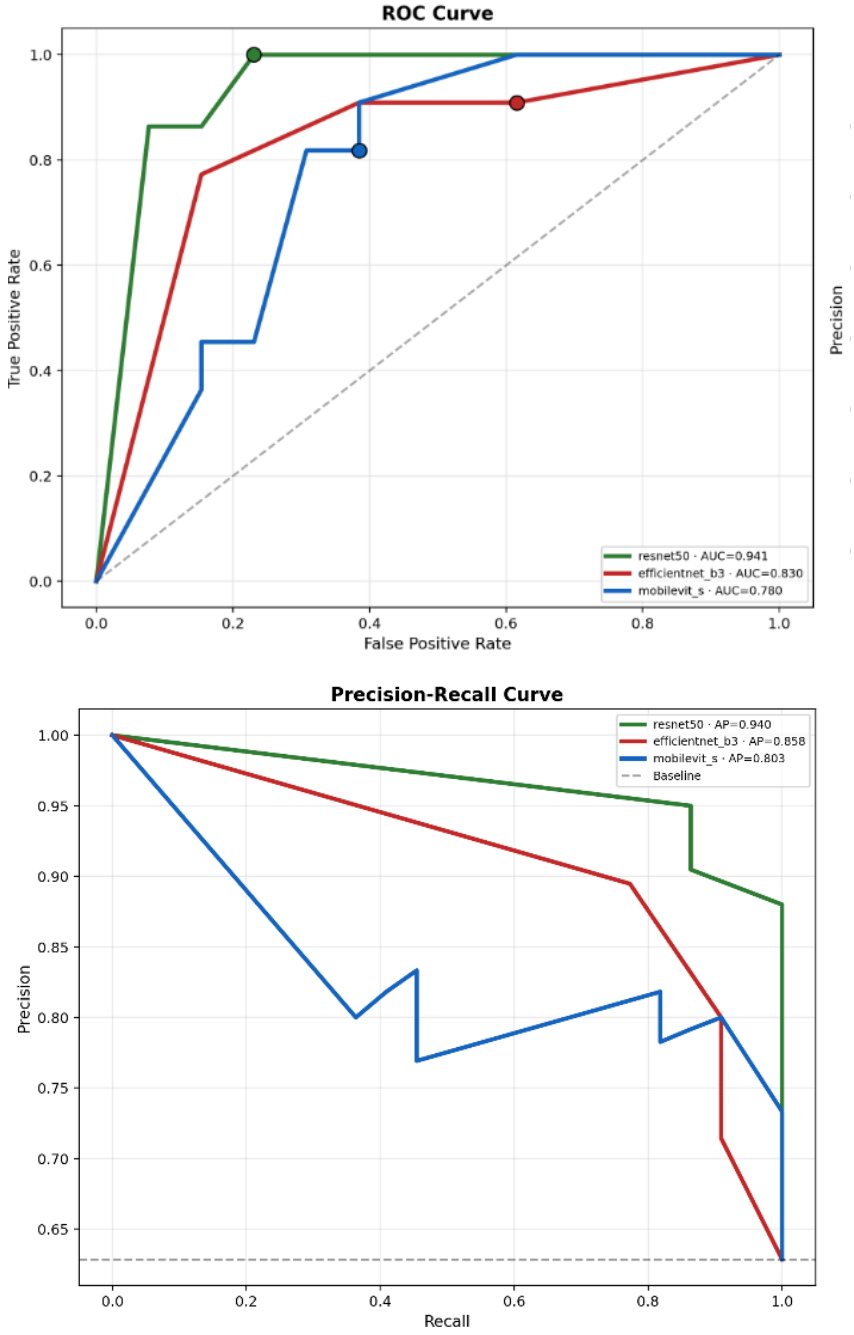


Tabla 5

Consolidado por experimento de curvas ROC y Precision - Recall (Average Precision)

Experimento	Modelo	ROC-AUC	PR-AUC (AP)
EXP 1	ResNet50	0.925	0.919
	EfficientNet-B3	0.806	0.843
	MobileViT-S	0.790	0.806
EXP 2	ResNet50	0.906	0.899
	EfficientNet-B3	0.827	0.858
	MobileViT-S	0.836	0.849
EXP 3	ResNet50	0.899	0.895
	EfficientNet-B3	0.806	0.843
	MobileViT-S	0.836	0.849
EXP 4	ResNet50	0.941	0.940
	EfficientNet-B3	0.830	0.858
	MobileViT-S	0.780	0.803
EXP 5	ResNet50	0.897	0.892
	EfficientNet-B3	0.827	0.858
	MobileViT-S	0.836	0.849

En resumen, estos resultados dejan claro que ResNet50 se adaptó muchísimo mejor a las imágenes y al cambio del contraste que los otros modelos. Al lograr esos números tan altos en el experimento cuatro, la red demuestra que caso no comete errores graves como para dejar pasar un tumor maligno (falsos negativos), lo que convierte a este flujo de trabajo en una herramienta muy segura y recomendada para ayudar a los médicos brindando una herramienta de apoyo para la toma de decisiones, que le permita priorizar los casos más urgentes en los hospitales.

6. 3 Interpretabilidad clínica mediante Grad-CAM

Con el fin de analizar la interpretabilidad de las arquitecturas evaluadas, se implementó la técnica Grad-CAM sobre imágenes del conjunto de prueba. Esta técnica permitió visualizar las regiones anatómicas que influenciaron las decisiones de clasificación de cada modelo, facilitando la evaluación clínica de las activaciones generadas durante la inferencia.

A diferencia de las métricas tradicionales de desempeño, Grad - CAM permitió observar si las arquitecturas realmente estaban centrando su atención sobre las regiones mamarias relevantes o sí, por el contrario, dependían de patrones difusos o información contextual no relacionada directamente con la lesión. Esto resultó especialmente importante en un entorno clínico, donde la interpretabilidad visual representa un componente fundamental para aumentar la confianza sobre los sistemas de apoyo diagnóstico basados en aprendizaje profundo.

Los análisis realizados mostraron diferencias importantes entre las arquitecturas evaluadas. ResNet50 presentó activaciones más focalizadas y anatómicamente más consistentes, EfficientNet-B3 mostró patrones más compactos y estables, mientras que MobileViT-S tendió a generar activaciones más amplias y difusas, priorizando información contextual global sobre localización anatómica específica.

El análisis comparativo entre experimentos permitió identificar una evolución progresiva entre estabilidad visual y coherencia clínica entre las tres arquitecturas. En los primeros experimentos se observaron variaciones menos estables y con mayor dispersión térmica, especialmente en MobileViT-S, mientras que en los experimentos posteriores, éstos mostraron una mejor delimitación anatómica y mejor consistencia en las predicciones.

ResNet50 mantuvo el comportamiento interpretativo más sólido a lo largo de los cinco experimentos, sus mapas de activación mostraron regiones focalizadas sobre áreas sospechosas, en

especial en casos malignos, manteniendo una alta precisión espacial y adecuada separación entre el tejido relevante y fondo anatómico. Este comportamiento es derivado también a sus altos valores de ROC - AUC y Precision - Recall observados anteriormente.

Tabla 6
Consolidado Grad- CAM correspondiente a los 5 experimentos.

Experimento	Arquitectura	Casos Correctos	Interpretabilidad	Localización Anatómica	Patrón de Atención	Evaluación Global
EXP 1	**ResNet50**	4/4	Alta	Alta	Activaciones focalizadas y estructuradas	Excelente coherencia clínica y alta precisión espacial
	EfficientNet-B3	N/A	No eval.	No eval.	Grad-CAM no generado correctamente	No fue posible realizar evaluación interpretativa
	MobileViT-S	3/4	Media	Baja	Activaciones difusas y distribuidas globalmente	Buena sensibilidad contextual, pero baja precisión anatómica
EXP 2	**ResNet50**	4/4	Alta	Alta	Activaciones compactas y mejor delimitación anatómica	Mayor estabilidad visual y excelente coherencia clínica
	EfficientNet-B3	3/4	Med-Alta	Media	Activaciones compactas y selectivas	Mejora importante en interpretabilidad y estabilidad
	MobileViT-S	3/4	Media	Baj-Med	Atención contextual amplia y parcialmente difusa	Mejor coherencia espacial respecto al EXP1
EXP 3	**ResNet50**	4/4	Alta	Alta	Activaciones focales con ligera separación periférica	Mantiene buena coherencia clínica con menor estabilidad
	EfficientNet-B3	3/4	Med-Alta	Media	Patrones conservadores y compactos	Comportamiento estable, aunque menos expresivo
	MobileViT-S	3/4	Media	Baja	Atención global extensa y baja precisión espacial	Menor interpretabilidad clínica
EXP 4	**ResNet50**	4/4	Muy Alta	Muy Alta	Activaciones altamente focalizadas y coherentes	Mejor comportamiento visual global del estudio
	EfficientNet-B3	4/4	Med-Alta	Media	Activaciones compactas y estables	Buen equilibrio entre interpretabilidad y estabilidad
	MobileViT-S	3/4	Media	Baj-Med	Atención contextual más estable	Mejor estabilidad global respecto a experimentos previos
EXP 5	**ResNet50**	4/4	Muy Alta	Muy Alta	Alta focalización con ligera dispersión periférica	Excelente discriminación visual y coherencia anatómica
	EfficientNet-B3	4/4	Med-Alta	Media	Activaciones compactas y consistentes	Modelo visualmente estable y conservado
	MobileViT-S	3/4	Media	Baj-Med	Persistencia de activaciones contextua	Menor precisión anatómica, pero adecuada sensibilidad

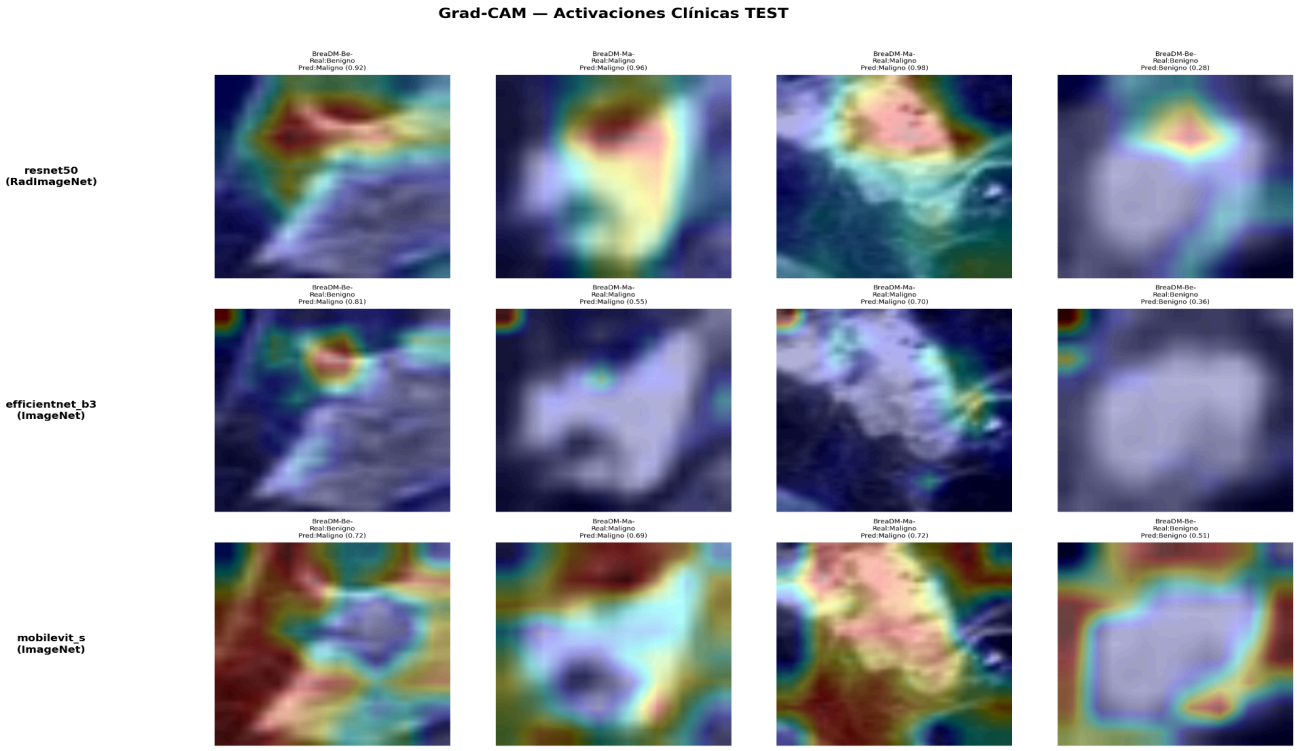
EfficientNet-B3, por su lado, presentó activaciones más compactas y conservadoras. Aunque sus mapas Grad - CAM mostraron menor intensidad visual en todos los experimentos y menor área abarcada de foco, ésta se centraba en áreas sospechosas, especialmente en zonas malignas, manteniendo así alta coherencia espacial y adecuada separación entre el tejido que es relevante.

Por último, MobileViT-S mostró mapas de activación difusos y amplios, abarcando así casi toda la imagen. Estas activaciones se vieron distribuidas sobre regiones contextuales de la mama en lugar de centrarse únicamente sobre la lesión principal, Aunque el modelo logró identificar regiones relacionadas con el área tumoral, presentó menor precisión anatómica e interpretabilidad clínica respecto de las demás arquitecturas.

6. 3.1 Visualización del mejor experimento

Figura 13

Grad- CAM correspondiente al experimento cuatro.



El experimento cuatro, fue seleccionado como el mejor experimento global, debido a que logró el equilibrio perfecto entre todas las métricas, aportando buen desempeño, estabilidad de entrenamiento, buena calibración probabilística e interpretabilidad clínica.

En este experimento, ResNet50 presentó el mejor comportamiento visual. Sus mapas de calor (Grad - CAM) mostraron regiones de una forma preciosa, junto encima del tumor o sobre áreas clínicamente sospechosas, sin distraerse con el tejido sano de alrededor, Además, esa precisión la mantuvo tanto en los casos benignos como en los malignos, lo que demuestra que es un modelo muy confiable.

EfficienteNet-B3 tampoco se quedó atrás, ya que mostró enfoques bien concentrados y limpios, con muy poco ruido en la imagen. Aunque no cubría tanta área del tumor como ResNet50, visualmente fue un modelo muy equilibrado y fácil de entender,

Por último, MobileViT-S siguió con su tendencia de mirar la imagen de manera más global y difusa, Aunque no es tan exacta apuntando el centro de la lesión, por lo menos logró entender el contexto general de la zona, sobre todo cuándo se trataba de tumores malignos.

Al final, el experimento cuatro ha dejado claro que en medicina no basta con que una inteligencia artificial tenga un puntaje perfecto, también es obligatorio que nos pueda demostrar exactamente en qué se está fijando para que el personal de salud especialista pueda confiar y apoyarse en ella.

6. 4 Comparación global entre arquitecturas

Con el fin de realizar una evaluación del desempeño experimental, se compararon las arquitecturas RestNet50, EfficientNe B3 y MobileViT-S considerando métricas clínicas, discriminativas, cuantitativas, además de factores relacionados con estabilidad de entrenamiento, capacidad de generalización, interpretabilidad clínica mediante GradCAM y costo computacional. En términos generales, los resultados que obtuve, evidenciaron que cada arquitectura aprendió de manera diferente el comportamiento temporal del contraste presente en las imágenes DCE-MRI. Aunque los tres modelos lograron clasificar lesiones benignas y malignas con desempeños competitivos, cada uno presentó fortalezas y limitaciones particulares relacionadas con sensibilidad diagnóstica, estabilidad probabilística, precisión anatómica y comportamiento clínico frente a los falsos negativos.

ResNet50 fue la arquitectura que mostró el comportamiento más equilibrado a lo largo de los

experimentos. Aunque en algunos escenarios presentó ligeras reducciones en sensibilidad respecto a otros modelos, logró mantener un desempeño robusto en métricas clínicas generales como AUC, F1-Score y Specificity. Además, presentó uno de los comportamientos más estables durante el entrenamiento, con mejor degradación de las curvas de validación y una mejor relación entre capacidad discriminativa y generalización clínica.

Por otro lado, EfficientNet-B3 mostró un comportamiento altamente sensible frente a los cambios de configuración experimental. En varios experimentos, logró mejorar significativamente el Recall y reducir la tasa de falsos negativos (FNR), aspecto importante dentro de un contexto clínico oncológico. Sin embargo, estas mejoras estuvieron acompañadas de reducciones importantes en Specificity, generando una mayor cantidad de falsos positivos y evidenciando una menor estabilidad frente a cambios de hiperparámetros.

En el caso de MobileViT-S, la arquitectura híbrida basada en mecanismos de atención, presentó resultados intermedios entre las CNN tradicionales y las arquitecturas más profundas. Aunque logró mantener tiempos de ejecución relativamente bajos y un comportamiento relativamente estable durante el entrenamiento, clínicamente mostró dificultades para sostener altos niveles de sensibilidad diagnóstica, ésto debido a que su infraestructura se ve limitada por datasets pequeños, esto fue una identificación importante dentro del entrenamiento.

Otro aspecto importante observado durante el análisis comparativo, fue el comportamiento de las curvas de entrenamiento. Algunas arquitecturas, especialmente ResNet50, mostraron tendencias claras de sobreajuste a partir de la época 12, esto fue evidenciado a partir de incrementos de la pérdida de validación mientras la pérdida de entrenamiento continuaba disminuyendo. Este comportamiento permitió identificar la importancia del uso de estrategias como early stopping, calibración

probabilística y ajuste de umbrales clínicos durante el flujo de trabajo experimental.

Desde el punto de vista clínico, uno de los factores más significativos dentro de esta investigación fue la reducción de falsos negativos, debido a que un falso negativo dentro del contexto de cáncer de mama, implica clasificar incorrectamente una lesión maligna como benigna. Por esta razón, durante el análisis experimental se priorizó el comportamiento de Recall y la tasa FNR por encima de métricas tradicionales como Accuracy.

Al final, los resultados obtenidos indicaron que cada red aprendió a leer los cambios del contraste de las resonancias a su manera. Aunque las tres arquitecturas hicieron un buen trabajo logrando un buen desempeño y clasificando los tumores entre benignos y malignos, cada modelo sacó a relucir sus propias fortalezas y debilidades en los experimentos, respecto con la sensibilidad diagnóstica, estabilidad probabilística, precisión anatómica e inferencia computacional.

La tabla 7 presenta una comparación global de las arquitecturas evaluadas durante el experimento cuatro, incluyendo métricas clínicas, probabilísticas y computacionales, generadas a través de MLFlow y que se encuentran en el repositorio en su tabla de presentación original para cada experimento.

Tabla 7

Tabla comparativa global correspondiente al Experimento cuatro

Métrica / Parámetro	resnet50	mobilevit_s	efficientnet_b3
Best Epoch	17	19	23
Best Phase	finetune	finetune	finetune
Train Auc	0.9996	0.9464	0.9221
Val Auc Patient	0.9545	0.9196	0.9476
Val Threshold	0.55	0.495	0.48
TEST AUC (cal)	0.8969	0.8357	0.8269
TEST AUC CI95	[0.766, 0.992]	[0.670, 0.964]	[0.678, 0.962]
TEST Recall	0.8636	0.7727	0.9091
TEST Specificity	0.8462	0.6923	0.5385
TEST F1	0.8837	0.7907	0.8333
TEST PPV	0.9048	0.8095	0.7692
TEST NPV	0.7857	0.6429	0.7778
TEST FNR	0.1364	0.2273	0.0909
TEST Brier	0.1048	0.1519	0.183
TP / TN / FP / FN	19/11/2/3	17/9/4/5	20/7/6/2
Recall ≥ 0.95	No	No	No
Calibrated (Isotónica)	Sí	Sí	Sí
Exec Time (Min)	1.67	1.61	2.66

6.5 Discusión general de resultados

Los resultados obtenidos durante el desarrollo experimental, permitieron evidenciar que las arquitecturas de aprendizaje profundo sí lograron aprender patrones importantes dentro de las imágenes DCE-MRI usadas en esta investigación. A través de las diferentes configuraciones implementadas, fue posible observar cómo cada modelo respondía de manera distinta frente a los cambios de hiperparámetros, estrategias de entrenamiento y ajustes clínicos realizados durante el flujo de trabajo experimental.

Aunque las tres arquitecturas lograron desempeños competitivos en la clasificacipn binaria de lesiones benignas y malignas, también quedó en eficiencia que no existe un modelo “perfecto” dentro

de este tipo de problemas médicos. Algunos modelos priorizaron sensibilidad diagnóstica, mientras otros lograron mejor especificidad o estabilidad durante el entrenamiento.

Además, el desarrollo iterativo de los experimentos permitió comprender que en problemas clínicos reales, no basta únicamente con obtener un valor alto de AUC, sino que también resulta importante analizar el comportamiento clínico completo del modelo especialmente en métricas relacionadas con falsos negativos y sensibilidad diagnóstica.

6. 5.1 Hallazgos principales

Uno de los principales hallazgos de esta investigación fue comprobar que la representación multicanal construida usando las fases PRE, EARLY y LATE del estudio DCE-MRI, permitió que las arquitecturas aprendieran información temporal importante relacionada con el comportamiento del contraste dentro de las lesiones mamarias.

También se observó que ResNet50 fue la arquitectura más estable y equilibrada a lo largo de todos los experimentos realizados, siempre era la mejor, manteniendo así buenos resultados clínicos incluso cuándo modificaba algunas configuraciones de entrenamiento e hiperparámetros. Por otro lado, EfficientNet-B3 mostró una mayor capacidad para aumentar sensibilidad diagnóstica, aunque sacriicó especificidad en algunos escenarios, al inal, MovbileViT-S logró resultados aceptables, pero presentó mayor dificultad para mantener estabilidad clínica frente a las otras arquitecturas evaluadas.

Otro hallazgo que fue importante para poder identificar que pequeños cambios experimentales podían general alteración y variaciones muy importantes en métricas clínicas como Recall y FNR. Esto permitió comprender que el desarrollo de modelos médicos requiere procesos de ajuste mucho más cuidadosos que en tareas tradicionales de clasificación.

6. 5.2 Impacto clínico de FN y Recall

Dentro del contexto clínico del cáncer de mama, uno de los aspectos más importantes durante esta investigación fue el análisis de falsos negativos (FN) y sensibilidad diagnóstica (Recall).
Un falso negativo representa un caso maligno clasificado incorrectamente como benigno, situación que puede generar retrasos en el diagnóstico, seguimiento y tratamiento oportuno de la paciente. Debido a esto, durante el desarrollo de los experimentos, se priorizó siempre el comportamiento del Recall sobre otras métricas tradicionales.

A lo largo de los experimentos fue posible observar cómo algunas configuraciones permitirían reducir falsos negativos, aunque en ocasiones esto también en ocasiones esto también afectaba notoriamente la especificidad, generando un aumento en la cantidad de falsos positivos y un desequilibrio en el diagnóstico. Este comportamiento mostró la importancia de lograr un equilibrio clínico entre la sensibilidad y especificidad dentro de sistemas de apoyo diagnóstico asistidos por inteligencia artificial.

6. 5.3 Limitaciones experimentales

A pesar de los resultados obtenidos, el desarrollo experimental presentó diversas limitaciones asociadas principalmente al tamaño del conjunto de datos, la capacidad computacional disponible y la complejidad computacional disponible al inicio del proyecto y la robustez que introducía complejidad de la infraestructura experimental usada.

Una de las principales limitaciones correspondió al tamaño del dataset empleado. Aunque la base de datos BreastDM contiene estudios clínicos reales de imágenes de resonancia magnética, la cantidad total de pacientes continúa siendo reducida para el entrenamiento de arquitecturas profundas

basadas en aprendizaje profundo y mecanismos de atención, Esta limitación puede afectar la capacidad de generalización de los modelos y aumentar el riesgo de sobreajuste de los modelos durante el entrenamiento.

Adicionalmente, el desarrollo experimental estuvo condicionado por restricciones de capacidad computacional. Durante distintas etapas del proyecto, se presentaron limitaciones relacionadas con memoria GPU, tiempos prolongados de entrenamiento y estabilidad en la ejecución de experimentos más robustos y complejos. Estas dificultades se hicieron más evidentes conforme el flujo experimental incorporó validación cruzada K-fold, generación de métricas, almacenamiento de archivos CSV, registros de ejecución, PKLs y mecanismos adicionales de análisis experimental, generando tiempos de entrenamiento superiores a una hora con recursos de capacidad de cómputo avanzada, finalmente obtenidos como recursos de la universidad.

Así mismo, el crecimiento progresivo de la infraestructura experimental generó dificultades relacionadas con trazabilidad y estabilidad de algunos procesos, incluyendo errores de serialización de modelos, corrupción de archivos PKL, pérdida parcial de resultados y fallos durante procesos de evaluación posteriores al entrenamiento. Cómo solución debido al limitante de tiempo disponible para estabilizar completamente el junto experimental, fue necesario reconstruir parcialmente la infraestructura y remover algunas técnicas implementadas.

Otra limitación importante estuvo relacionada con el alcance inicialmente planteado para la infraestructura experimental. Algunas estrategias avanzadas y mecanismos complementarios preparados en fases anteriores tuvieron que ser simplificados o descartados debido a restricciones de tiempo y estabilidad experimental, priorizando la entrega de un flujo de trabajo estable, reproducible, y metodológicamente consistente que garantizara al menos 5 experimentos realizados.

A pesar de estas limitaciones, fue posible consolidar una infraestructura experimental funcional

orientada al entrenamiento, validación y comparación de diferentes arquitecturas de aprendizaje profundo aplicadas a clasificación binaria de lesiones mamarias mediante imágenes DCE-MRI.

6. 5.4 Potencial clínico futuro

A pesar de tratarse de una investigación desarrollada en un contexto académico de pregrado, los resultados obtenidos permitieron evidenciar el potencial que posee las arquitecturas de aprendizaje profundo como herramientas de apoyo para el análisis de imágenes médicas relacionadas con cáncer de mama.

El uso de representaciones multicanal DCE-MRI permitió que los modelos aprendieran patrones asociados al comportamiento dinámico del contraste dentro del tejido mamario, aspecto clínicamente importante para la diferenciacipón entre lesiones benignas y malignas. Esto demuestra que la integración de inteligencia artificial dentro del análisis radiológico puede convertirse en una herramienta complementaria para apoyar procesos de diagnóstico asistido.

Adicionalmente, el uso de métricas clínicas, calibración probabilística y herramientas de interpretabilidad visual como GradCAM, permitió desarrollar un enfoque más cercano a escenarios clínicos reales, dónde no solamente importa la capacidad predictiva del modelo, sino también, la comprensión de su comportamiento y el anślisis de posibles errores diagnósticos.

Como trabajo futuro, esta investigación podrá ser extendida mediante evaluación multicéntrica con bases de datos más grandes y diversas, que incorporen no sólo la clasificación binaria de DCE-MRI, sino también una clasificación de cáncer en su fase de diagnóstico final qué es durante la fase posterior a la resonancia magnética, la biopsia de mama, esto integrando un trabajo académico

también sólido que he venido realizando en la base de datos de breat cáncer de Wisconsin. También me gustaría empezar a trabajar segmentación de diferentes tipos de cáncer de mama y unificar como propuesta la realización de un CAD para consolidar las fases de diagnóstico de cáncer mamario.

7. Conclusiones

El desarrollo de esta investigación permitió evidenciar que las arquitecturas de aprendizaje profundo aplicadas sobre representaciones multicanal DCE-MRI poseen un alto potencial como herramientas de apoyo para la clasificación de lesiones mamarias benignas y malignas.

A través de la implementación de un flujo de trabajo experimental basado en transferencia de aprendizaje, ajuste fino progresivo y evaluación clínica a nivel de paciente, fue posible entrenar y comparar diferentes arquitecturas de aprendizaje profundo, incluyendo ResNet50, EfficientNet-B3 y MobileViT-S, obteniendo comportamientos diferenciados entre modelos tanto a nivel computacional como clínico.

Los resultados obtenidos demostraron que ResNet50 fue la arquitectura más robusta y estable dentro del contexto experimental desarrollado, alcanzando los mejores resultados globales en métricas clínicas como AUC, Recall y FNR. Esto permitió evidenciar que arquitecturas convolucionales profundas continúan mostrando un comportamiento altamente competitivo en tareas médicas con datasets relativamente pequeños y especializados.

Por otro lado, arquitecturas como EfficientNet-B3 y MobileViT-S permitieron analizar diferentes comportamientos relacionados con sensibilidad clínica, generalización y estabilidad experimental. En particular, EfficientNet-B3 mostró mejoras importantes en Recall durante algunos experimentos, aunque sacrificando Specificity y aumentando falsos positivos, mientras que MobileViT-S evidenció ciertas limitaciones de estabilidad posiblemente asociadas al tamaño reducido

del dataset y a la naturaleza híbrida Transformer-CNN de la arquitectura.

La utilización de representaciones multicanal derivadas de fases PRE, EARLY y LATE permitió conservar información temporal clínicamente relevante relacionada con la dinámica de captación del contraste dentro del tejido mamario, favoreciendo el aprendizaje de patrones asociados a lesiones malignas.

Adicionalmente, la incorporación de métricas clínicas, calibración probabilística y herramientas de interpretabilidad visual como GradCAM permitió desarrollar un enfoque más cercano a escenarios clínicos reales, donde no solamente importa la capacidad predictiva del modelo, sino también la comprensión de su comportamiento y la identificación de posibles errores diagnósticos.

Finalmente, esta investigación permitió consolidar conocimientos relacionados con aprendizaje profundo aplicado a imágenes médicas, procesamiento DCE-MRI, reproducibilidad experimental y evaluación clínica de modelos de inteligencia artificial, dejando además una base experimental sólida para futuras investigaciones orientadas al desarrollo de sistemas CAD más robustos y clínicamente aplicables.

8. Recomendaciones

Como trabajo futuro, se recomienda continuar esta línea de investigación utilizando bases de datos multicéntricas más grandes y diversas, permitiendo validar el comportamiento de las arquitecturas bajo escenarios clínicos más heterogéneos y cercanos a entornos hospitalarios reales.

También se recomienda explorar arquitecturas híbridas más avanzadas basadas en mecanismos de atención y Vision Transformers, especialmente en contextos donde se disponga de una mayor cantidad de datos médicos para entrenamiento, ya que este tipo de modelos suele requerir volúmenes de información más amplios para alcanzar un comportamiento estable.

Adicionalmente, podría resultar interesante incorporar tareas de segmentación tumoral y localización automática de lesiones mamarias, permitiendo complementar la clasificación binaria realizada en esta investigación con análisis más detallados sobre la morfología y extensión de los tumores.

Otra recomendación importante consiste en integrar información clínica complementaria, como resultados histopatológicos, datos demográficos o antecedentes médicos, con el objetivo de construir modelos multimodales capaces de mejorar la capacidad diagnóstica y reducir posibles ambigüedades presentes únicamente en las imágenes DCE-MRI.

Finalmente, se recomienda continuar fortaleciendo estrategias relacionadas con interpretabilidad, calibración probabilística y control de reproducibilidad experimental, debido a que estos aspectos resultan fundamentales para aumentar la confianza clínica y facilitar futuras aplicaciones de herramientas de inteligencia artificial dentro de escenarios médicos reales.

9. Bibliografía.

Bonilla-Sepúlveda, Ó. A. (2022). Retrasos en el diagnóstico y tratamiento del cáncer de mama en Medellín, Colombia. *Ginecología y obstetricia de México*, *90*(12), 943-950. https://doi.org/10.24245/gom.v90i12.8114

Breast MRI Findings: Kinetics - Radiology | UCLA Health. (s. f.). Recuperado 22 de mayo de 2026, de https://www.uclahealth.org/departments/radiology/education/breast-imaging-teaching-resources/birads/breast-mri-findings-kinetics

Cheng, L., & Li, X. (2013). Breast magnetic resonance imaging: Kinetic curve assessment. *Gland Surgery*, *2*(1), 50-53. https://doi.org/10.3978/j.issn.2227-684X.2013.02.04

Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach. (2022). *Biomedical Signal Processing and Control*, *71*, 103141. https://doi.org/10.1016/j.bspc.2021.103141

Convolutional Neural Networks and Vision Transformers for Fashion MNIST Classification: A Literature Review. (s. f.). Recuperado 22 de mayo de 2026, de https://arxiv.org/html/2406.03478v1

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* (arXiv:2010.11929). arXiv. https://doi.org/10.48550/arXiv.2010.11929

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115-118. https://doi.org/10.1038/nature21056

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, ROC Analysis in*

Pattern Recognition, *27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Global Cancer Observatory. (s. f.). Recuperado 22 de mayo de 2026, de https://gco.iarc.who.int/

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2021, enero 4). *Transformers in Vision: A Survey*. arXiv.Org. https://doi.org/10.1145/3505244

Kim, J. W., Khan, A. U., & Banerjee, I. (2025). Systematic Review of Hybrid Vision Transformer Architectures for Radiological Image Analysis. *Journal of Imaging Informatics in Medicine*, *38*(5), 3248-3262. https://doi.org/10.1007/s10278-024-01322-4

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60-88. https://doi.org/10.1016/j.media.2017.07.005

Mehta, S., & Rastegari, M. (2022). *MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer* (arXiv:2110.02178). arXiv. https://doi.org/10.48550/arXiv.2110.02178

Sheth, D., & Giger, M. (2019). Artificial intelligence in the interpretation of breast cancer on MRI. *Journal of Magnetic Resonance Imaging*, *51*, 1310-1324. https://doi.org/10.1002/jmri.26878

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, junio 12). *Attention Is All You Need*. arXiv.Org. https://arxiv.org/abs/1706.03762v7

Zhao, X., Liao, Y., Xie, J., He, X., Zhang, S., Wang, G., Fang, J., Lu, H., & Yu, J. (2023a). BreastDM: A DCE-MRI dataset for breast tumor image segmentation and classification. *Computers in Biology and Medicine*, *164*, 107255. https://doi.org/10.1016/j.compbiomed.2023.107255

Zhao, X., Liao, Y., Xie, J., He, X., Zhang, S., Wang, G., Fang, J., Lu, H., & Yu, J. (2023b). BreastDM: A DCE-MRI dataset for breast tumor image segmentation and classification. *Computers in Biology and Medicine*, *164*, 107255. https://doi.org/10.1016/j.compbiomed.2023.107255

10. Anexos

Anexo A. Repositorio del proyecto y código fuente desarrollado en GitHub
Link:
https://github.com/PandoraRiot/MRI_BreastCancer_Classification/blob/main/experimentos/experimento5/figuras/gradcam_outputs/fig_gradcam_clinical_final.png

Anexo B. Base de datos pública BreastDM (ROI) utilizada durante el desarrollo experimental.
Link: https://drive.google.com/file/d/1rKps09z1DEkfiICZlpIBOFek-LZ1vFAC/view