Metodología para clasificación BI-RADS en mamografías

Laboratorio de Procesamiento de Imágenes Digitales, Universidad Nacional de Costa Rica

Abstract

We present a methodological framework for automated BI-RADS categorization in mammography that prioritizes rigor, scalability, and clinical applicability over hyperparameter optimization. The approach consists of: (i) object detection to isolate the breast region using a modern YOLO family detector, enabling automatic dataset cleaning and standardized cropping; (ii) cloud-based workflows to ensure reproducibility, collaboration, and auditable experiment tracking; (iii) principled data reorganization and inspection to address class imbalance across ordinal BI-RADS categories; (iv) strong data augmentation to emulate clinical variability; and (v) two-phase training with transfer learning—first adapting the classification head, then gradual fine-tuning—using efficient convolutional architectures inspired by the MobileNet philosophy (accuracy–efficiency trade-off suitable for constrained settings). Evaluation emphasizes per-class metrics, confusion matrices, and ordinal-aware agreement measures, coupled with model versioning and single-case testing to reflect practical usage. Rather than reporting peak scores, we articulate a generalizable method aligned with emerging reporting guidelines for AI in medical imaging. This framework is intended as a reproducible baseline for prospective, multi-center validation and for integration into screening workflows, especially in resource-limited environments.

Keywords: Breast cancer screening; Mammography; BI-RADS; Medical imaging; Artificial intelligence; Deep learning; Clinical decision support; Reproducible research; Data quality; Global health.

1. Introducción

El cáncer de mama es la principal causa de mortalidad oncológica en mujeres, con aproximadamente 2.3 millones de casos en 2020 [1]. La mamografía es el pilar del tamizaje, y su interpretación se estructura mediante BI-RADS, que asigna categorías ordinales vinculadas con riesgo de malignidad y conductas clínicas [2]. Pese a esta estandarización, persiste una variabilidad interobservador relevante —sobre todo en BI-RADS 3 y 4— que impacta la precisión diagnóstica y la eficiencia de los programas de cribado, agravada por el creciente volumen de estudios [3].

En este escenario, el aprendizaje profundo ha mostrado rendimientos comparables a los de radiólogos en detección de cáncer de mama [4]–[6]. Sin embargo, la evidencia se ha concentrado en configuraciones binarias (maligno/benigno), mientras que la clasificación multiclase y de naturaleza ordinal de BI-RADS sigue siendo un desafío que exige rigor en la definición de la unidad de análisis, control del desbalance de clases y evaluación alineada con la práctica clínica [7].

Con base en esta necesidad, este trabajo presenta una metodología integral y reproducible para la clasificación BI-RADS que prioriza solidez y transparencia por encima del ajuste fino de hiper parámetros. La propuesta articula cinco decisiones clave: (i) depuración anatómica automática para estandarizar el campo de visión y eliminar contenido no diagnóstico; (ii) entorno de trabajo en la nube para trazabilidad, colaboración y control de versiones; (iii) reorganización e inspección rigurosa del conjunto de datos, con especial atención al desbalance ordinal y a la unidad de análisis; (iv) aumentación motivada clínicamente para emular variabilidad de adquisición y posicionamiento; y (v) aprendizaje por transferencia en dos fases —adaptación del clasificador y ajuste fino gradual con tasas reducidas—. El objetivo es ofrecer una base replicable y transferible a la práctica, en sintonía con guías contemporáneas de reporte en inteligencia artificial aplicada a imagen médica.

2. Metodología

La metodología propuesta se estructura como una secuencia coherente y reproducible que transforma las imágenes crudas en una evaluación de relevancia clínica. Cada etapa delimita con precisión su propósito, entradas y salidas, y fija controles mínimos que aseguran la consistencia y la comparabilidad entre ejecuciones. En lo que sigue, se expone el flujo metodológico con un equilibrio adecuado entre detalle procedimental y claridad expositiva, preservando la continuidad analítica de principio a fin.

Paso 1 - Estandarización de la entrada (detección y recorte).

El proceso inicia con la localización automática de la región mamaria mediante un detector de la familia YOLO sin restricción de clase (umbral de confianza ajustable). A partir de esa detección se recorta el contenido estrictamente diagnóstico, eliminando bordes, rótulos y artefactos. Sustituir reglas manuales por detección aprendida aporta consistencia ante variaciones de adquisición y entre centros, y reduce sesgos derivados de información no clínica. La implementación admite procesamiento en lote: recorre subdirectorios, procesa múltiples imágenes en segundos y guarda cada recorte. El resultado es un conjunto de imágenes homogéneas y listas para las etapas siguientes.

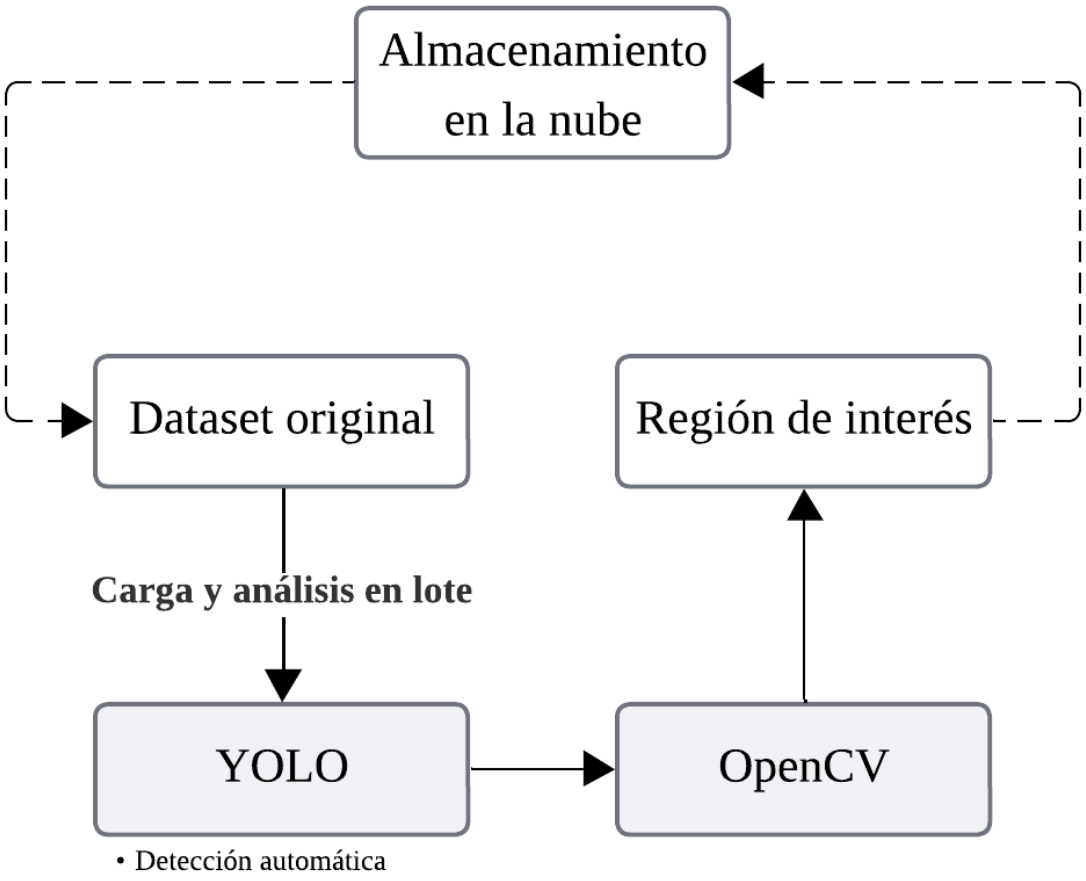

Figura 1. Estandarización de entrada con detección y recorte

Paso 2 - Ejecución en la nube (reproducibilidad y colaboración).

Sobre esa base, el trabajo se ejecuta en un entorno de nube con el fin de garantizar reproducibilidad extremo a extremo y facilitar la colaboración. Datos, código y resultados quedan versionados; las corridas registran semillas y configuraciones; y se aprovechan aceleradores (GPU/TPU) según disponibilidad. Este esquema reduce la variabilidad entre corridas y permite comparar cambios de forma ordenada, sin depender del hardware local.

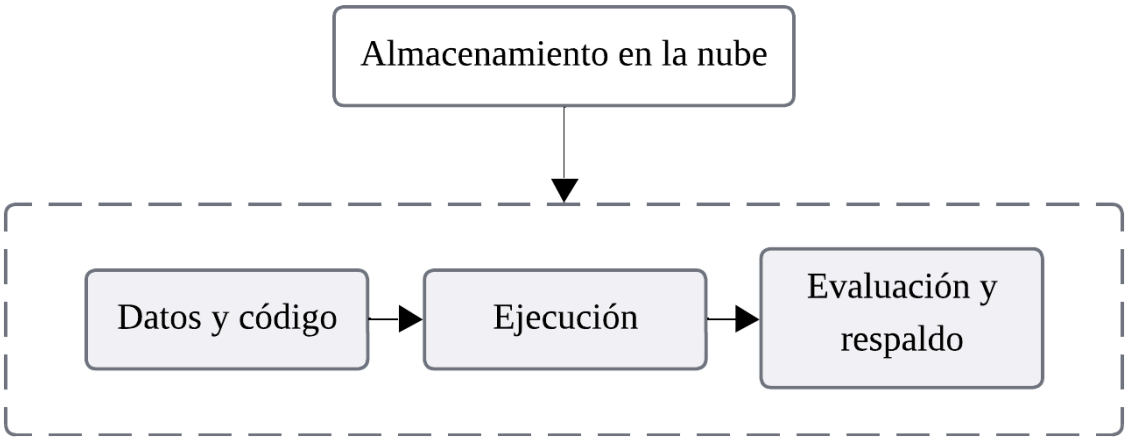

Figura 2. Entorno en la nube: datos y código → ejecución orquestada → artefactos (pesos, curvas, respaldos).

Paso 3 - Reorganización e inspección del conjunto de datos.

Con los recortes estandarizados, se reorganiza el conjunto en particiones de entrenamiento, validación y prueba, definiendo con claridad la unidad de análisis (estudio, mama o hallazgo) y evitando fugas entre particiones (por ejemplo, que imágenes de una misma paciente aparecen en conjuntos diferentes). Cuando el tamaño muestral es limitado, la validación cruzada estratificada resulta preferible a una única división. La inspección cuantitativa (conteos por categoría BI-RADS, distribución por vistas CC/MLO, detección de duplicados) caracteriza el desbalance habitual —abundancia de 1–2 frente a escasez de 4–5— y orienta las decisiones posteriores.

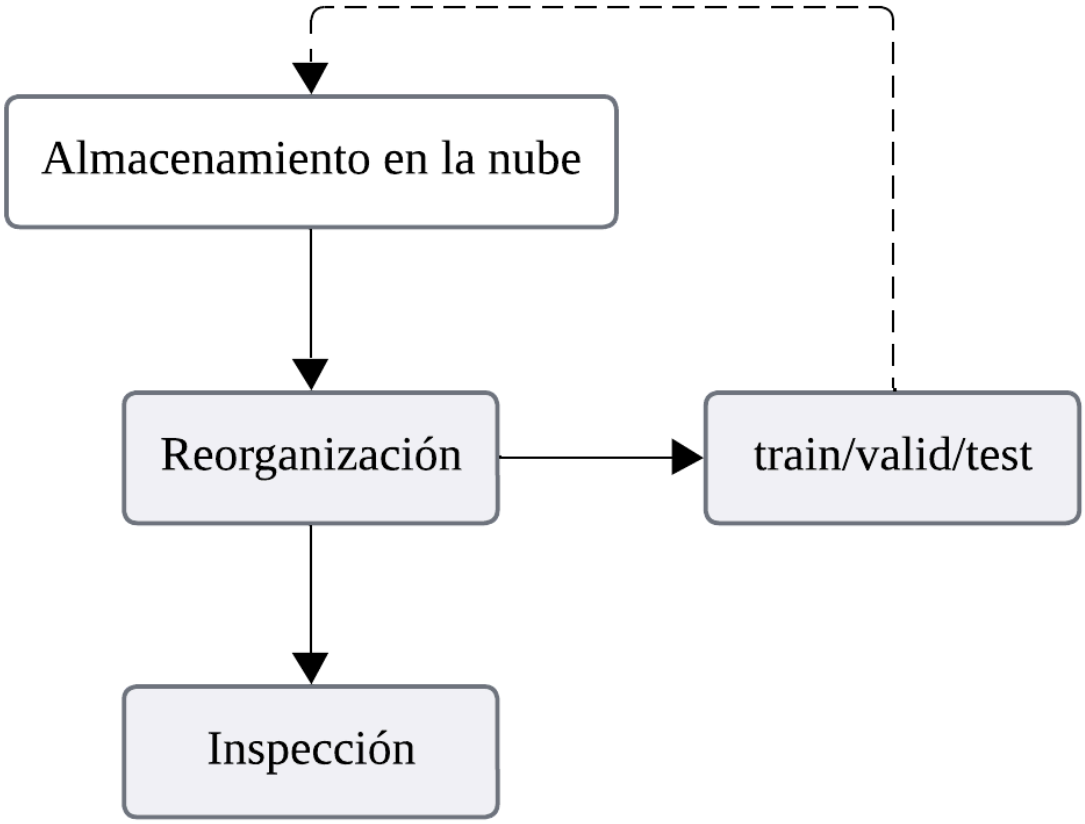

Figura 3. Reorganización en entrenamiento/validación/prueba y chequeo rápido de distribución por clase.

Paso 4 - Aumentación con sentido clínico y preparación de entrada.

La aumentación busca inducir invariancias plausibles sin alterar la anatomía: rotaciones y traslaciones para simular variaciones de posicionamiento; zoom, recorte y redimensionado (p. ej., a 224×224) para robustez espacial; y ajustes moderados de brillo/contraste junto con perturbaciones controladas para reflejar diferencias de adquisición. En paralelo, se optimiza el flujo de datos con caché, prelectura y barajado reproducible, de modo que el acelerador trabaje sin esperas y las comparaciones entre ejecuciones sean justas.

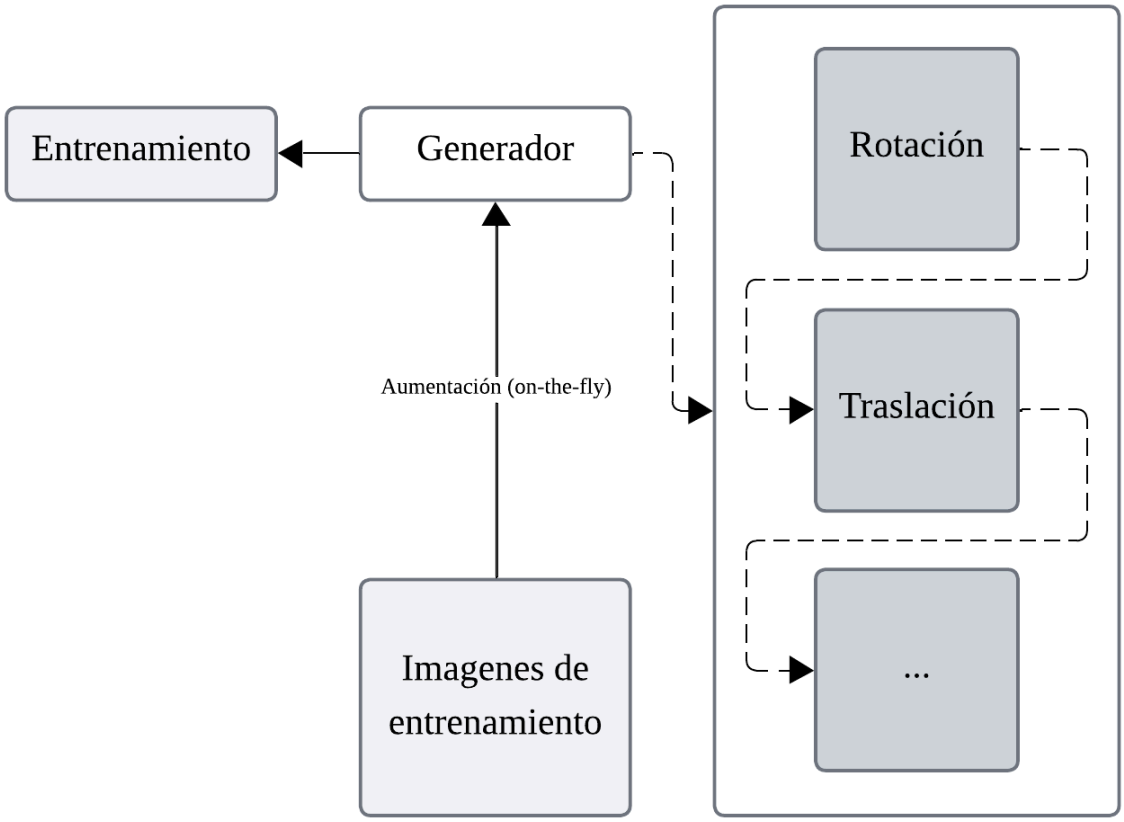

Figura 4. Cadena de aumentación y preparación de entrada.

Paso 5 - Entrenamiento en dos fases con transferencia.

El aprendizaje se estructura en dos etapas. Primero, se adapta únicamente la cabeza de clasificación mientras la base convolucional pre entrenada permanece congelada; así se alinea la salida con la tarea BI-RADS sin perturbar representaciones útiles. Después, se descongelan gradualmente las últimas capas y se realiza un ajuste fino con tasa de aprendizaje menor y parada temprana, conservando el mejor punto según validación. Esta secuencia estabiliza la convergencia, reduce el sobreajuste en conjuntos modestos y mantiene el equilibrio entre desempeño y costo computacional.

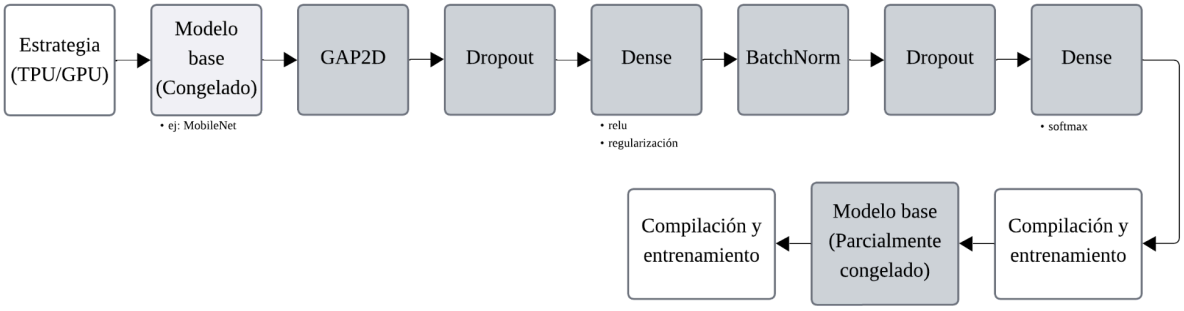

Figura 5. Dos fases: (F1) cabeza sobre base congelada → (F2) ajuste fino parcial con tasa de aprendizaje reducida.

Paso 6 - Evaluación alineada con el uso clínico.

La evaluación refleja la naturaleza ordenada de BI-RADS. Además de la exactitud global, se reportan métricas por clase y matrices de confusión; cuando corresponde, se incluyen medidas sensibles al orden y calibración de probabilidades, útiles para tareas como priorización o segunda lectura. Los modelos se guardan con sello temporal, se conservan casos fallidos para análisis de error y, como verificación práctica, se realiza la prueba de una imagen individual, que reproduce el uso en una estación de lectura clínica.

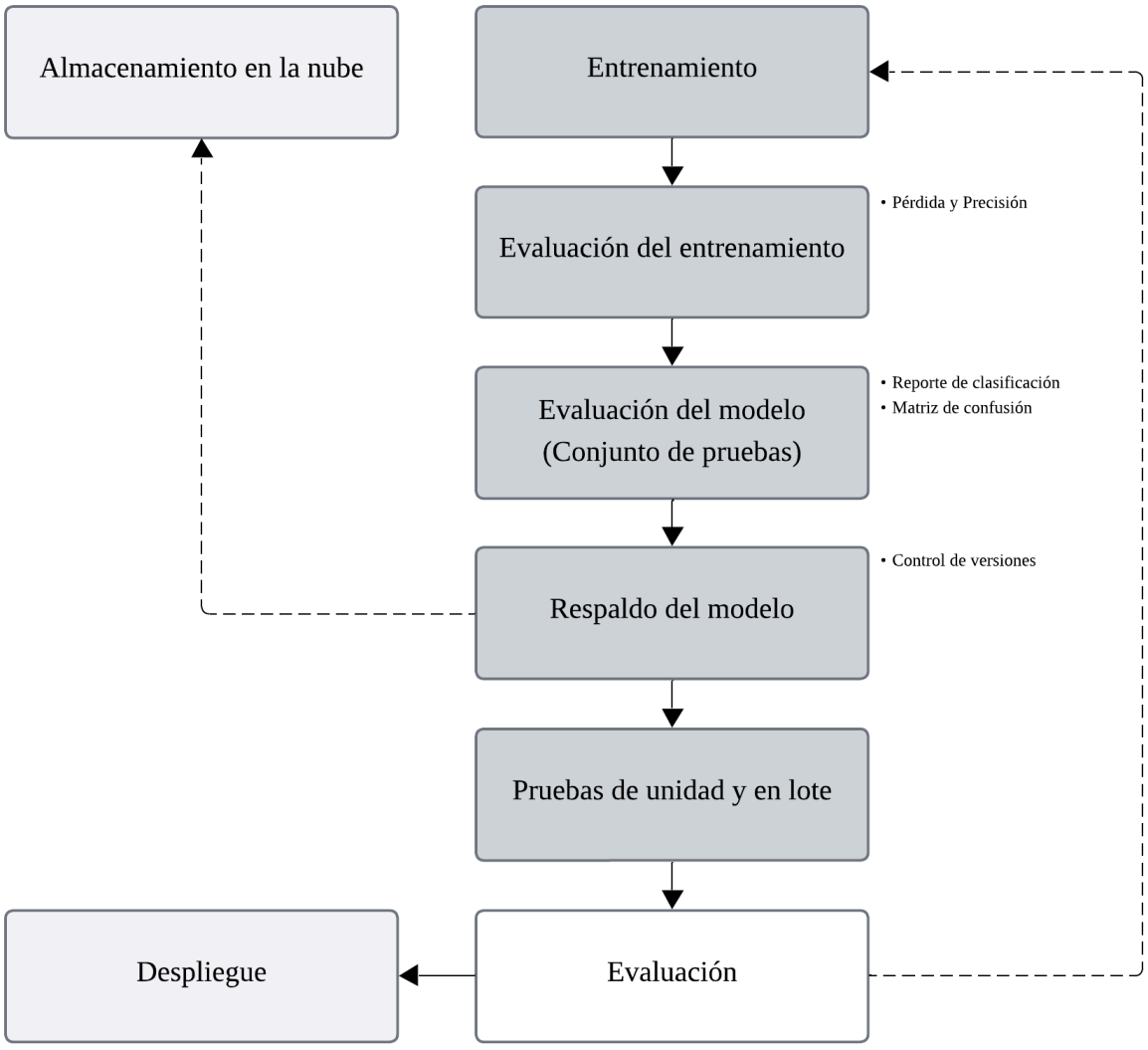

Figura 6. Evaluación del modelo y entrenamiento.

En conjunto, este flujo define un camino claro y reproducible: estandarizar lo que entra, controlar lo que se ejecuta, entender lo que se entrena y evaluar de acuerdo con la tarea real. Esa continuidad metodológica es la que habilita comparaciones honestas y una eventual transferencia a entornos clínicos.

3. Evaluación del método

La separación del aprendizaje en dos fases produjo curvas de entrenamiento más estables y un beneficio consistente del ajuste fino en validación. Tal comportamiento es coherente con la hipótesis de adaptación progresiva: primero se adecua la cabeza de clasificación al dominio y luego se especializan las capas finales sin destruir las representaciones útiles previas. Las matrices de confusión evidenciaron, como es esperable, confusiones entre categorías intermedias (2 con 3, 3 con 4), un fenómeno que también se observa en la práctica clínica. La incorporación

de medidas macro y ordinales evitó conclusiones engañosas basadas en promedios; por su parte, la calibración post entrenamiento mejoró la confiabilidad de las probabilidades, lo cual es crucial cuando la salida del sistema se emplea para priorización o como segunda lectura.
Desde la perspectiva operativa, la localización anatómica automática y la reorganización reproducible del conjunto de datos redujeron tiempos y errores manuales, y facilitaron la trazabilidad experimental. La eficiencia del flujo de datos permitió aprovechar de forma sostenida los aceleradores, reduciendo la variabilidad entre ejecuciones. En conjunto, estos elementos configuran una metodología preparada para su evaluación multicéntrica y su eventual integración en flujos de tamizaje.

4. Discusión

El aporte central de este trabajo es metodológico. La localización anatómica aprendida reemplaza procedimientos frágiles de preprocesamiento y homogeneiza la entrada del clasificador, reduciendo la varianza de presentación entre estudios. La aumentación con plausibilidad clínica, combinada con transferencia y ajuste fino gradual, ofrece un camino prudente y reproducible en escenarios con datos limitados, mitigando el sobreajuste y preservando rasgos relevantes. De forma complementaria, la orientación a la reproducibilidad —versionado de experimentos, fichas de datos y de modelo, y documentación consistente— fortalece la confiabilidad del pipeline y facilita su transferencia a la práctica.

Persisten, no obstante, desafíos clave. La definición de etiquetas y de la unidad de análisis (imagen, mama, paciente, episodio) condiciona qué se predice y cómo se evalúa, por lo que requiere protocolización explícita. La naturaleza ordinal de BI-RADS demanda métricas acordes (p. ej., acuerdo ponderado) y calibración; además, los análisis por subgrupos son necesarios para detectar sesgos y caracterizar equidad. Ninguna arquitectura garantiza, por sí sola, transportabilidad: su utilidad debe examinarse bajo restricciones reales de latencia, memoria y consumo energético. Finalmente, la validez externa depende de evaluaciones prospectivas y multicéntricas que midan no sólo desempeño diagnóstico, sino también impacto operativo (tiempos de lectura, necesidad de segunda lectura, tasas de llamada) y que mantengan supervisión humana continua.

En conjunto, estos resultados sugieren que el valor no reside en un conjunto específico de hiper parámetros, sino en principios operativos verificables: estandarizar lo que entra, controlar lo que se ejecuta, comprender lo que se entrena y evaluar según el uso clínico real. Sin estos elementos, las mejoras algorítmicas difícilmente se traducen en beneficios clínicos tangibles.

6. Trabajo futuro

Para orientar la agenda de investigación y su eventual transferencia a la práctica clínica, se esbozan las siguientes líneas de trabajo, en carácter de propuesta:

1. **Validación prospectiva y multicéntrica**: Sería valioso emprender estudios prospectivos en múltiples centros que contemplen métricas diagnósticas y operativas —incluidos tiempos de lectura, segunda lectura y tasas de llamada—, además de ensayos de integración en flujos reales y sistemas PACS.

2. **Evaluación ampliada como estándar**: Convendría incorporar de forma sistemática la estimación de incertidumbre, la explicabilidad visual y el análisis de equidad por subgrupos como componentes regulares del protocolo de evaluación.

3. **Eficiencia y escalabilidad bajo restricciones reales**: Resultaría útil comparar arquitecturas eficientes y escalables bajo límites realistas de latencia, memoria y energía; asimismo, podría analizarse el escalado de la resolución de entrada y su impacto en eficiencia, desempeño y calibración.

4. **Pre Entrenamiento específico del dominio**: Podría explorarse el pre entrenamiento auto-supervisado para reducir la dependencia de etiquetas y mejorar la adaptabilidad a distintos entornos clínicos.

Idealmente, estas líneas deberían acompañarse de protocolos transparentes, reportes reproducibles y artefactos de trazabilidad que faciliten la comparación entre centros y una adopción responsable.

7. Conclusiones

En síntesis, este trabajo presenta un marco metodológico coherente y reproducible, orientado a la práctica clínica, para la clasificación BI-RADS en mamografías. La propuesta integra estandarización anatómica, un entorno de ejecución en la nube, reorganización e inspección rigurosa de datos, aumentación guiada por criterios clínicos y aprendizaje por transferencia con ajuste fino gradual. La evaluación se alinea con la naturaleza ordinal de la tarea e incorpora calibración de probabilidades; junto con prácticas de trazabilidad y gobernanza, constituye una base realista para estudios prospectivos y para una adopción responsable en contextos con recursos limitados. Más que enumerar hiper parámetros, el trabajo consolida principios y decisiones justificadas que habilitan la replicación, la comparación y la transferencia clínica.

Referencias

[1] American College of Radiology, BI-RADS Atlas, 5th ed., 2013.

[2] S. M. McKinney et al., International evaluation of an AI system for breast cancer screening, Nature, 2020.

[3] A. Yala et al., A Deep Learning Model to Triage Screening Mammograms, Radiology, 2019.

[4] A. Rodríguez-Ruiz et al., Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography, JNCI, 2019.

[5] I. C. Moreira et al., INbreast: Toward a Full-Field Digital Mammographic Database, Academic Radiology, 2012.

[6] The Cancer Imaging Archive (TCIA), CBIS-DDSM, 2017.

[7] H. T. Nguyen et al., VinDr-Mammo, 2022–2023.

[8] M. Sandler et al., MobileNetV2, CVPR, 2018; A. Howard et al., MobileNetV3, ICCV, 2019.

[9] M. Tan and Q. V. Le, EfficientNet: Rethinking Model Scaling, ICML, 2019.

[10] M. Raghu et al., Transfusion: Understanding Transfer Learning for Medical Imaging, NeurIPS, 2019.

[11] J. Mongan et al., Checklist for Artificial Intelligence in Medical Imaging (CLAIM), Radiology: AI, 2020.

[12] G. S. Collins et al., TRIPOD+AI Statement, BMJ, 2024.

[13] X. Liu et al., CONSORT-AI; S. C. Rivera et al., SPIRIT-AI, BMJ, 2020.

[14] C. Guo et al., On Calibration of Modern Neural Networks, ICML, 2017.