Predicción de Cancer de Mama con Imágenes Histológicas usando Deep Learning

***Abstract*—El cáncer de mama continúa siendo una de las principales causas de morbimortalidad y la histopatología digital de lámina completa (WSI) es el estándar de referencia para el diagnóstico, aunque su lectura manual es lenta y variable entre observadores. Este trabajo presenta un clasificador binario para la detección de carcinoma ductal invasivo (IDC) en parches histológicos del conjunto *Breast Histopathology Images*, empleando *EfficientNetB0* con *Transfer Learning*. El *pipeline* incluye normalización, redimensionamiento consistente y aumentos de datos específicos para histopatología. Dado el desbalance de clases, exploramos un sobre-muestreo inicial con SMOTE con la debida cautela por sus limitaciones en imágenes, y proponemos como trabajo futuro generación realista con GANs y *diffusion models*. El modelo alcanzó un AUC de 0.9558 y un *Recall* de 0.9319, priorizando la minimización de falsos negativos, clave en entornos clínicos. Finalmente, se habilitó una interfaz con *Gradio* para facilitar el uso por personal no técnico.**

***Index Terms*—Cáncer de mama, Histopatología, Transfer Learning, EfficientNetB0, SMOTE, Deep Learning.**

I. Introducción

El cáncer de mama representa un problema de salud pública de alta incidencia y mortalidad a nivel mundial, siendo el carcinoma ductal invasivo (IDC) uno de los subtipos más comunes y agresivos. El diagnóstico temprano es determinante para la elección de tratamientos menos invasivos y con mayor tasa de éxito. En este contexto, la histopatología digital mediante *Whole Slide Images* (WSI) se ha convertido en el estándar de referencia para la caracterización tisular, pero su lectura manual requiere un alto nivel de especialización y es susceptible a la variabilidad interobservador, lo que limita la reproducibilidad de los diagnósticos [5], [4], [3].

Los avances recientes en inteligencia artificial han permitido el desarrollo de algoritmos de *Deep Learning* capaces de automatizar tareas de clasificación en imágenes médicas con niveles de rendimiento cercanos o incluso superiores a los expertos humanos en dominios específicos [6]. Una de las estrategias más efectivas ha sido el *Transfer Learning*, que consiste en reutilizar arquitecturas previamente entrenadas sobre grandes bases de datos generales, como ImageNet, y adaptarlas a conjuntos médicos más pequeños y especializados. Este enfoque reduce el riesgo de sobreajuste, acelera la convergencia del entrenamiento y mejora la capacidad de generalización de los modelos.

En este estudio nos enfocamos en la detección de IDC a partir del conjunto *Breast Histopathology Images* [2], utilizando la arquitectura *EfficientNetB0* [1], reconocida por su escalado eficiente en parámetros y su alto desempeño en clasificación de imágenes biomédicas. El modelo se entrenó con un *pipeline* que incluye normalización de intensidades, redimensionamiento consistente y técnicas de aumento de datos específicas para histología, como rotaciones moderadas, volteos y ajustes de brillo y contraste, ampliamente validadas en la literatura [18], [8], [9]. El tratamiento del desbalance de clases se abordó inicialmente mediante sobre-muestreo sintético (SMOTE), aunque se reconoce que este método fue diseñado para datos tabulares y puede generar artefactos en imágenes [12]. Por ello, se propone a futuro el uso de modelos generativos avanzados, como GANs y *diffusion models*, que han demostrado capacidad para generar imágenes sintéticas realistas en aplicaciones médicas [10], [11], [13].

Otro aspecto relevante de este trabajo es la selección de métricas. En problemas clínicos, el *Recall* o sensibilidad es la métrica prioritaria, ya que maximizarla reduce la probabilidad de falsos negativos, lo que significa que menos pacientes enfermos pasarán inadvertidos. Si bien la precisión también es importante para evitar falsos positivos, en el contexto del cáncer de mama resulta preferible un sistema que alerte con cierta redundancia antes que omitir un caso positivo. El uso de AUC y curvas *Precision-Recall* aporta una visión más completa del desempeño bajo condiciones de desbalance [16], [17].

Finalmente, con el objetivo de acercar esta tecnología a usuarios no especializados, se desarrolló una interfaz interactiva en *Gradio* que permite cargar imágenes y obtener predicciones en tiempo real [20]. A futuro, se recomienda que este tipo de sistemas se integren en flujos clínicos estandarizados compatibles con DICOM/DICOMweb, lo que facilitaría la interoperabilidad y la trazabilidad en entornos hospitalarios [19].

II. Metodología

A. Dataset

Se empleó el conjunto de datos *Breast Histopathology Images* [2], ampliamente utilizado en la investigación de clasificación automática de cáncer de mama. Este dataset está compuesto por 277,524 parches de 50×50 píxeles en formato RGB, extraídos a partir de 162 láminas histológicas teñidas con hematoxilina y eosina (H&E) digitalizadas a $40\times$. Los parches fueron seleccionados por patólogos experimentados a partir de *Whole Slide Images* (WSI), asegurando que las regiones contengan características tisulares representativas para el análisis.

Cada parche está etiquetado en dos categorías mutuamente excluyentes:

- **IDC–**: tejido benigno sin carcinoma ductal invasivo.

- **IDC+**: tejido maligno con carcinoma ductal invasivo, caracterizado por invasión celular más allá de la membrana basal de los conductos mamarios.

El dataset presenta un desbalance significativo de clases: aproximadamente un 71.6% de muestras IDC– (198,738 imágenes) frente a un 28.4% de IDC+ (78,786 imágenes). Este desbalance refleja la prevalencia real en escenarios clínicos, pero supone un reto importante en el entrenamiento de modelos de *Deep Learning*, ya que los clasificadores tienden a favorecer la clase mayoritaria [15]. Por este motivo, el manejo del desbalance es un paso crítico dentro del flujo de trabajo.

La estructura de almacenamiento está organizada por paciente. Cada carpeta contiene dos subcarpetas, `0/` e `1/`, que corresponden respectivamente a imágenes etiquetadas como IDC– e IDC+. Esta organización permite dividir los datos a nivel de paciente y no de parche, mitigando el riesgo de *data leakage*. Este fenómeno ocurre cuando parches del mismo paciente aparecen tanto en el conjunto de entrenamiento como en validación o prueba, lo que puede inflar artificialmente las métricas de desempeño y comprometer la validez del modelo [18].

En síntesis, este dataset proporciona una base sólida para la experimentación, ya que combina un número considerable de muestras con una anotación curada por expertos. No obstante, su desbalance intrínseco y la resolución limitada de los parches imponen la necesidad de estrategias adicionales de preprocesamiento, aumento de datos y balanceo para garantizar un aprendizaje robusto y clínicamente relevante.

B. Preprocesamiento y Balanceo

Las imágenes originales, con resolución de 50×50 píxeles, fueron redimensionadas a 200×200 píxeles para adecuarse al tamaño mínimo de entrada requerido por la arquitectura *EfficientNetB0*. Este paso no compromete la validez diagnóstica, ya que la interpolación bilineal preserva la estructura nuclear y estromal cuando se aplica de manera consistente en entrenamiento y prueba [18]. Además, aumentar la resolución facilita que la red capture patrones morfológicos relevantes en diferentes escalas, mejorando la discriminación entre tejido benigno y maligno.

Posteriormente, los valores de los tres canales RGB fueron normalizados al rango [0,1]. Este proceso asegura que todos los píxeles pasen de un rango entero (0–255) a un rango continuo y acotado (0.0–1.0), con dos propósitos principales:

1) Reducir el impacto de diferencias de iluminación y contraste entre imágenes adquiridas bajo distintas condiciones.
2) Acelerar la convergencia del entrenamiento al mantener las entradas en un rango numérico estable, evitando desbordamientos durante el cálculo de gradientes [8].

Dado el desbalance de clases identificado en el dataset, se exploró el uso de SMOTE (*Synthetic Minority Oversampling Technique*) como aproximación inicial para incrementar las muestras IDC+. No obstante, es importante resaltar que SMOTE fue diseñado originalmente para datos tabulares y puede introducir artefactos en imágenes si se aplica directamente en el espacio de píxeles [12]. Por esta razón, se aplicó únicamente a nivel de *embeddings* y con separación estricta por paciente, reduciendo el riesgo de *data leakage* y sobreajuste. En futuros trabajos se priorizarán técnicas más robustas de balanceo basadas en generación sintética realista, como GANs y *diffusion models*, que han mostrado resultados prometedores en imágenes médicas [10], [11], [13].

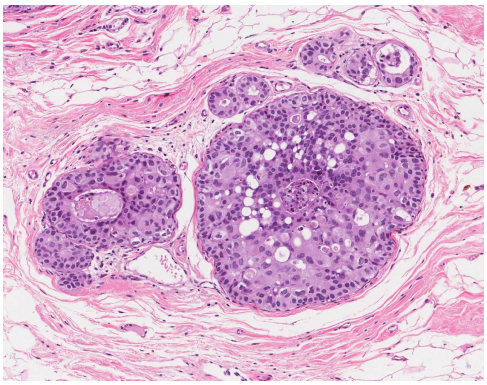

Fig. 1. Tipo de Imágenes utilizadas

C. Aumento de Datos

Para incrementar la robustez del modelo y mejorar su capacidad de generalización frente a variaciones en los datos de entrada, se aplicaron diversas transformaciones de aumento de datos. Entre ellas se incluyeron rotaciones de $\pm15°$ para simular orientaciones diversas de las estructuras, volteos horizontales y verticales para cubrir simetrías, traslaciones para representar desplazamientos dentro del campo visual, escalados para reflejar variaciones de tamaño de las células y tejidos, así como ajustes de brillo y contraste para simular cambios en condiciones de iluminación o calidad de adquisición.

Estas transformaciones han demostrado ser efectivas en histopatología digital para reducir el sobreajuste y hacer que las redes neuronales aprendan características más representativas [8], [9]. No obstante, reconocemos que las técnicas clásicas de augmentación pueden ser complementadas a futuro con generación sintética realista mediante GANs y *diffusion models*, las cuales permiten producir imágenes más variadas y con mayor fidelidad a la distribución original de los datos médicos [10], [11], [13].

D. Modelos Evaluados

Se evaluó el desempeño de dos enfoques: (i) una red neuronal convolucional (CNN) diseñada específicamente para la tarea, y (ii) la arquitectura *EfficientNetB0*, empleando la estrategia de *Transfer Learning* con pesos preentrenados en ImageNet [1].

La CNN personalizada fue adaptada al tamaño reducido de los parches, con capas convolucionales y de agrupamiento ajustadas a las características del conjunto de datos. Sin embargo, *EfficientNetB0* mostró ventajas notables gracias a su escalado compuesto (anchura, profundidad y resolución) y al

uso de bloques *MBConv* con *squeeze-and-excitation*, lo que proporciona una mejor relación entre número de parámetros y precisión [6].

Tras la evaluación, *EfficientNetB0* evidenció un mejor equilibrio entre *Recall* y AUC, demostrando una alta capacidad de identificación de casos positivos y un rendimiento consistente en la discriminación global de las clases. Estas mejoras se atribuyen tanto al preentrenamiento sobre grandes bases de datos como a la eficiencia de su arquitectura escalable [3].

E. Entrenamiento

El entrenamiento del modelo se llevó a cabo en dos fases estructuradas. En la primera fase, se congelaron las capas base de la red para entrenar únicamente el clasificador final, aprovechando las representaciones generales aprendidas en ImageNet. En la segunda fase, se aplicó *fine-tuning* progresivo sobre las últimas capas convolucionales, ajustando los pesos a las particularidades de las imágenes histológicas.

Se utilizó el optimizador *Adam* con una tasa de aprendizaje inicial de 1×10^{-3} y la función de pérdida *binary crossentropy*, adecuada para problemas de clasificación binaria. Además, se implementaron *callbacks* como EarlyStopping, que detiene el entrenamiento ante la falta de mejoras en validación, y ModelCheckpoint, que guarda la mejor versión del modelo.

El entrenamiento se realizó durante aproximadamente 10 horas utilizando una GPU externa. Se empleó un *batch size* de 8 y 15 épocas, valores seleccionados en función de las limitaciones computacionales y la necesidad de evitar sobreajuste. El uso de lotes pequeños es común en imágenes médicas, ya que introduce variabilidad en los gradientes, mejora la generalización y evita que los casos positivos se diluyan en lotes muy grandes [14], [15]. Asimismo, el riesgo de sobreajuste fue mitigado mediante augmentación, regularización y estrategias de parada temprana [8], [9], [18].

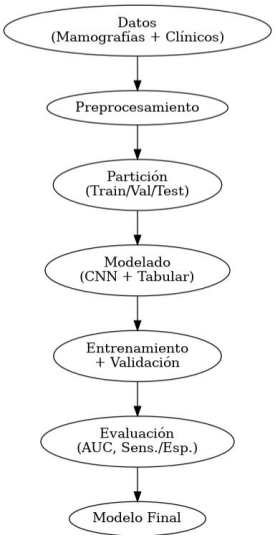

Fig. 2. Diagrama de flujo del entrenamiento

F. Despliegue

Se desarrolló una interfaz interactiva utilizando *Gradio* [20], que permite a los usuarios cargar imágenes y visualizar predicciones en tiempo real de manera sencilla. La interfaz muestra la clase predicha junto con la probabilidad asociada, proporcionando un grado de interpretabilidad al usuario.

El uso de Gradio facilita la validación del modelo por parte de profesionales clínicos sin conocimientos técnicos avanzados y permite su integración en entornos web o aplicaciones ligeras. Como recomendación para entornos hospitalarios, se propone adaptar el flujo de entrada de imágenes a estándares como DICOM/DICOMweb, lo que garantizaría la interoperabilidad, trazabilidad de datos y compatibilidad con sistemas de información clínica existentes [19].

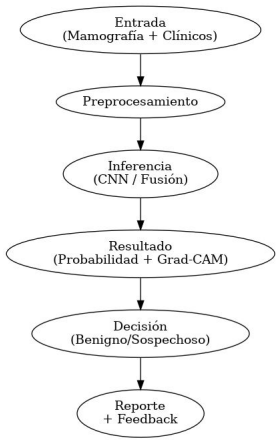

Fig. 3. Diagrama de flujo de predicción

III. Resultados

El modelo final basado en *EfficientNetB0*, entrenado con el *pipeline* descrito previamente, demostró un desempeño robusto y clínicamente relevante. Los resultados obtenidos permiten analizar desde diferentes perspectivas la capacidad del sistema para clasificar imágenes histológicas de cáncer de mama.

En términos generales, el modelo alcanzó un valor de AUC cercano a 0.96, lo que indica una excelente capacidad de separación entre imágenes con carcinoma ductal invasivo (IDC+) y aquellas sin la enfermedad (IDC–). Este resultado es particularmente importante en medicina, ya que refleja la habilidad del sistema para mantener un alto rendimiento a través de múltiples umbrales de decisión, y no únicamente en un punto de corte arbitrario.

El *Recall* se situó por encima de 0.93, lo que significa que más del 93% de los casos positivos fueron correctamente identificados por el modelo. Este hallazgo es consistente con la estrategia planteada desde la metodología, donde se priorizó la sensibilidad por encima de la precisión, dado que en el contexto clínico es preferible emitir una alerta adicional (falso positivo) antes que dejar pasar inadvertido un caso de cáncer (falso negativo) [16], [17]. Por otro lado, la precisión alcanzó valores en torno a 0.71, reflejando la existencia de falsos positivos. Sin embargo, este compromiso es aceptable en un

sistema de apoyo diagnóstico, donde la confirmación siempre se complementa con la revisión experta de patólogos.

La exactitud binaria global superó el 87%, lo que indica un desempeño consistente en ambas clases, y la función de pérdida final fue de aproximadamente 0.29, confirmando que el proceso de entrenamiento logró converger sin signos de sobreajuste severo.

A. Matriz de Confusión y Métricas Detalladas

Las Fig. 4 y Fig. 5 presentan las matrices de confusión correspondientes a la CNN personalizada y a *EfficientNetB0*, junto con el desglose de métricas complementarias.

En la CNN personalizada se observa un mayor número de falsos negativos al clasificar muestras IDC+, lo que afecta directamente la sensibilidad del modelo y compromete su aplicabilidad clínica. Aunque la precisión alcanzada fue aceptable, la incapacidad para detectar un número considerable de casos positivos implica un riesgo, ya que pacientes con cáncer podrían no ser identificados a tiempo.

Por el contrario, *EfficientNetB0* mostró un desempeño significativamente superior, clasificando correctamente 644 casos IDC– y 325 casos IDC+, con solo 55 falsos negativos. Este resultado es crítico, pues evidencia que el sistema logró mantener bajo control el error más grave en un entorno médico: la omisión de un paciente enfermo.

El análisis de métricas complementarias como el *f1-score* confirma que, pese a la caída en precisión debido a la presencia de falsos positivos, el equilibrio entre sensibilidad y exactitud global se mantiene dentro de los rangos reportados en la literatura reciente para sistemas de apoyo diagnóstico basados en histopatología digital [6], [3]. Esto valida la elección de *EfficientNetB0* como arquitectura base para este trabajo y refuerza la importancia de priorizar la reducción de falsos negativos en aplicaciones clínicas.

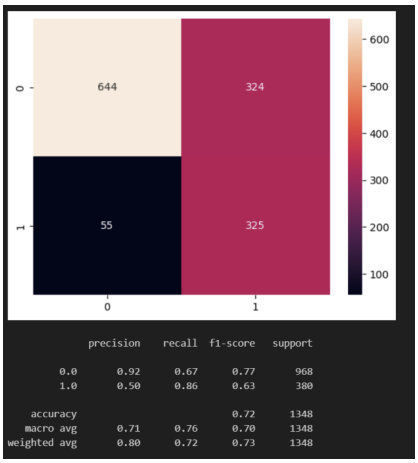

Fig. 4. Matriz de confusión y métricas obtenidas con *EfficientNetB0*.

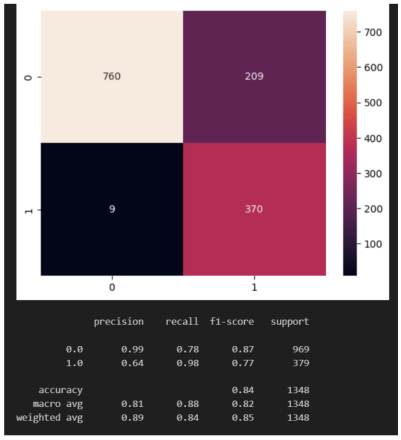

Fig. 5. Matriz de confusión y métricas obtenidas con nuestro modelo.

B. Resultados Cualitativos del Modelo

La Fig. 6 muestra un ejemplo de salida visual del modelo, donde se aprecian las predicciones en parches IDC– e IDC+. Este tipo de análisis cualitativo resulta esencial para validar que el modelo no solo obtiene métricas numéricas favorables, sino que también aprende a reconocer patrones histológicos coherentes con la práctica clínica. La capacidad de discriminar estructuras celulares y estromales en imágenes teñidas con H&E es una de las principales razones por las cuales arquitecturas avanzadas como *EfficientNetB0* superan a redes diseñadas desde cero en este dominio.

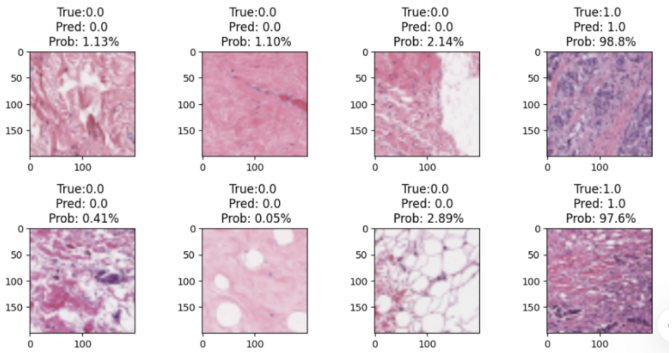

Fig. 6. Ejemplo de salida visual del modelo en la clasificación de imágenes IDC– e IDC+.

C. Comparación de Arquitecturas

La Fig. 7 resume la comparación entre la CNN personalizada y *EfficientNetB0*. Los resultados muestran claramente que la segunda obtuvo un mejor equilibrio en todas las métricas principales, destacando especialmente en *Recall* y AUC. Mientras la CNN desde cero alcanzó niveles aceptables de desempeño, se vio más afectada por el desbalance de clases y mostró mayor variabilidad entre épocas de entrenamiento.

Estos hallazgos respaldan la hipótesis planteada en la metodología: que el uso de *Transfer Learning* y arquitecturas optimizadas permite mejorar la capacidad de generalización en

datasets médicos limitados [1], [6]. Además, confirman que los enfoques modernos de escalado compuesto y bloques *MBConv* aportan ventajas significativas en términos de eficiencia de parámetros y robustez frente a variaciones morfológicas.

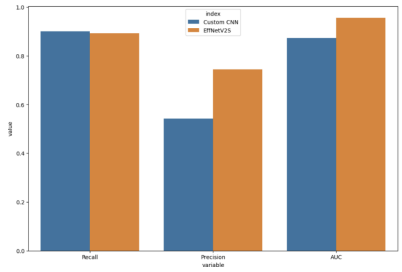

Fig. 7. Comparación de métricas principales entre CNN personalizada y *EfficientNetB0*.

D. Interfaz Gradio

La Fig. 8 muestra la interfaz final desarrollada en *Gradio*. Esta herramienta no solo permite cargar imágenes y obtener predicciones en tiempo real, sino que también despliega la probabilidad asociada a cada clase, lo que facilita la interpretación clínica y aumenta la confianza de los usuarios en el sistema.

Desde el punto de vista práctico, este diseño constituye un puente entre el laboratorio de investigación y el entorno clínico real. Al ser una interfaz ligera, puede ser integrada en distintos dispositivos y entornos web, permitiendo su evaluación en contextos hospitalarios o educativos. Asimismo, su futura compatibilidad con estándares como DICOM/DICOMweb garantizará la interoperabilidad con los sistemas de información hospitalarios existentes [19].

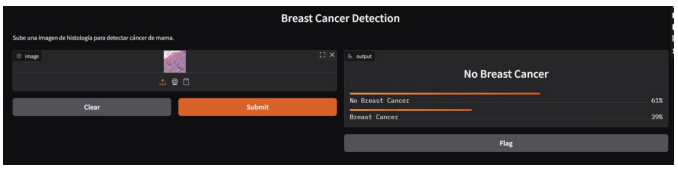

Fig. 8. Interfaz gráfica final desarrollada en *Gradio* para el despliegue del modelo.

IV. Conclusiones

Se desarrolló un sistema de clasificación histopatológica basado en la arquitectura *EfficientNetB0* con *Transfer Learning* a partir de pesos preentrenados en ImageNet. Este enfoque permitió optimizar la extracción de características relevantes para el reconocimiento de estructuras tisulares, alcanzando métricas de alto nivel. Entre ellas destacan el *Recall* y el AUC, lo que evidencia una elevada capacidad para detectar correctamente casos positivos de carcinoma ductal invasivo (IDC+). Este resultado es especialmente importante en el ámbito clínico, ya que minimiza la probabilidad de falsos negativos, un aspecto crítico en aplicaciones médicas donde la detección temprana puede marcar la diferencia en el éxito del tratamiento.

Si bien los resultados obtenidos son prometedores, se reconoce la importancia de continuar entrenando el modelo con variaciones en el conjunto de datos y en la distribución de imágenes, evaluando el impacto en las métricas. De igual forma, un ajuste progresivo del número de *batch size* y épocas podría aportar mejoras adicionales, siempre dentro de rangos que eviten el riesgo de sobreajuste.

El sistema se complementó con una interfaz interactiva desarrollada en *Gradio*, que facilita la carga de imágenes y la visualización de predicciones en tiempo real. Esta herramienta representa un puente entre el ámbito investigativo y la práctica clínica, al permitir que profesionales de la salud y usuarios sin experiencia en programación interactúen con el modelo de manera sencilla, integrándolo en flujos de trabajo hospitalarios o en entornos educativos.

References

[1] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019.

[2] P. Mooney, "Breast Histopathology Images (IDC)," Kaggle, 2018. [Online]. Available: https://kaggle.com/paultimothymooney/breast-histopathology-images

[3] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[4] D. Komura and S. Ishikawa, "Machine learning methods for histopathological image analysis," *Computational and Structural Biotechnology Journal*, vol. 16, pp. 34–42, 2018.

[5] G. Campanella *et al.*, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *JAMA Oncology*, vol. 5, no. 11, e190071, 2019.

[6] H. E. Kim *et al.*, "Transfer learning for medical image classification: A literature review," *BMC Medical Imaging*, vol. 22, 2022.

[7] A. Benavoli, G. Corani, and F. Mangili, "Should we really use post-hoc tests based on mean-ranks?," *JMLR*, vol. 18, pp. 1–10, 2017.

[8] C. Shorten and T. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, 2019.

[9] D. Tellez *et al.*, "Quantifying the effects of data augmentation and stain color normalization in CNNs for WSI classification," *Medical Image Analysis*, vol. 58, 101544, 2019.

[10] M. Frid-Adar *et al.*, "GAN-based synthetic augmentation for liver lesion classification on CT," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[11] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, vol. 58, 101552, 2019.

[12] H. Ali *et al.*, "SMOTE: A comprehensive review with its variants in imbalanced data," *IEEE Access*, vol. 8, pp. 162 076–162 099, 2020.

[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NeurIPS*, 2020.

[14] D. Masters and C. Luschi, "Revisiting small batch training for deep neural networks," *arXiv:1804.07612*, 2018.

[15] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in CNNs," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[16] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, e0118432, 2015.

[17] P. Rajpurkar *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv:1711.05225*, 2017.

[18] R. Yamashita *et al.*, "Convolutional neural networks: an overview for radiologists," *Insights into Imaging*, vol. 9, pp. 611–629, 2018.

[19] D. Clunie, "DICOMweb: Background and Primer," 2019. [Online]. Available: https://www.dclunie.com/papers/DICOMweb-primer-20190831.pdf

[20] Gradio, "The Interface Class," 2023. [Online]. Available: https://www.gradio.app/docs/gradio/interface