ISSN 1870-4069

Evaluación comparativa de la representatividad de modelos RNCP en mastografías públicas del Hospital General de Ensenada

J.I. Ayala-Guebara[1,2], J.A. González-Fraga[1], J. Magaña-Magaña[2], E. Gutiérrez-López[1], G. J. Avilés-Rodríguez[3], L.M. Pellegrin-Zazueta[1]

[1] Universidad Autónoma de Baja California,
Facultad de Ciencias,
México

[2] Hospital General Ensenada,
México

[3] Universidad Autónoma de Baja California,
Escuela de Ciencias de la Salud,
México

{angel_fraga, luis.pellegrin}@uabc.edu.mx

Resumen. Las redes neuronales convolucionales profundas (RNCP) son ampliamente utilizadas en el aprendizaje supervisado y han demostrado su capacidad para aprender relaciones entrada-salida en grandes volúmenes de imágenes. En este artículo se presenta un estudio comparativo de la representatividad de modelos de RNCP entrenados con 2 conjuntos de datos distintos. El primer modelo se genera utilizando el dataset público CBIS-DDSM, el cual clasifica las imágenes de mastografía en categorías de benigno y maligno. El segundo modelo se entrena con la base de datos de mastografías del Hospital General de Ensenada, que clasifica las imágenes según la escala BI-RADS del 1 al 5. El tercero se entrena y evalúa con ambas bases de datos. Se generaron varios modelos utilizando las arquitecturas MobileNetV2, VGG16, Resnet, CNN, DenseNet e Inception. Tras su evaluación MobileNetV2 demostró el mejor rendimiento. Este estudio analiza y compara la exactitud (accuracy), precisión, sensibilidad y f1-score de los modelos para evaluar la representatividad y efectividad de cada conjunto de datos en la clasificación de imágenes de mastografía. Los hallazgos de esta investigación buscan proporcionar información valiosa para el desarrollo y mejora de herramientas de diagnóstico asistido por computadora, con el objetivo de optimizar el diagnóstico temprano y el tratamiento del cáncer de mama en diferentes entornos clínicos.

Palabras clave: Cáncer de mama, mastografía, redes neuronales, convolucionales profundas, aprendizaje supervisado, MobileNetV2, diagnóstico asistido por computadora.

Comparative Evaluation of the Representativeness of Deep Convolutional Neural Network Models in Public Mammograms from the General Hospital of Ensenada

Abstract. Deep convolutional neural networks (DCNNs) are widely used in supervised learning and have demonstrated their ability to learn input-output relationships in large volumes of images. This paper presents a comparative study of the representativeness of DCNN models trained on two different datasets. The first model is generated using the public CBIS-DDSM dataset, which classifies mammogram images into benign and malignant categories. The second model is trained on the mammogram database from the General Hospital of Ensenada, which classifies images according to the BI-RADS scale from 1 to 5. The third model is trained and evaluated with both datasets. Several models were generated using the MobileNetV2, VGG16, ResNet, CNN, DenseNet, and Inception architectures. After evaluation MobileNetV2 demonstrated the best performance. This study analyzes and compares the accuracy, precision, sensitivity, and F1-score of the models to assess the representativeness and effectiveness of each dataset in mammogram image classification. The findings of this research aim to provide valuable insights for the development and improvement of computer-aided diagnostic tools, with the goal of optimizing early diagnosis and treatment of breast cancer in diverse clinical settings.

Keywords: Breast cancer, mammography, deep convolutional neural networks, supervised learning, MobileNetV2, computer-aided diagnosis.

1. Introducción

El cáncer de mama actualmente es un problema de salud pública que requiere ser atendido de manera prioritaria. De acuerdo a la Organización Mundial de la Salud (OMS), el cáncer es una de las principales causas de muerte en el mundo, se reporta que 1 de cada 6 muertes se debe al cáncer [1]. De acuerdo con datos del Instituto Nacional de Estadística y Geografía (INEGI), en el año 2020 fallecieron un total de 97 323 personas por tumores malignos; 7 880 fueron por tumores malignos de mama, lo que equivale al 8 % de este total. Debido al cáncer de mama fallecieron 7 821 mujeres y 58 hombres [2]. Se ha identificado el cáncer de mama como el tipo de neoplasia maligna con mayor incidencia y mortalidad en las mujeres a nivel global. En México el cáncer de mama es el tipo que más afecta a las mujeres, y está catalogado como la principal causa de muerte por este padecimiento. En el último reporte global de la Agencia Internacional para la Investigación del Cáncer, se reportaron más de 27 000 nuevos casos y casi 7 000 decesos debidos a esta causa [3]. Debido a esto, el cáncer de mama es considerado como un problema de salud pública nacional, cuya consecuencia es un promedio de más de 18 muertes por día. La detección temprana del cáncer de mama es crucial para aumentar las tasas de supervivencia y mejorar los resultados del tratamiento. Los modelos de Red Neuronal Convolucional Profunda (RNCP) han demostrado un gran potencial en la identificación de anomalías en mastografías [4]. Sin embargo, la representatividad de estos modelos es un factor crítico que determina su

eficacia en diversas poblaciones. Este estudio se centra en evaluar la representatividad de los modelos de RNCP utilizando 2 bases de datos: CBIS-DDSM [5] que es una base de datos pública y HGE-DB la cual es una base de datos privada bajo el resguardo del Hospital General de Ensenada.

Este documento se organiza en cinco secciones principales: Introducción, Trabajo relacionado, Metodología, Resultados, Discusión y Conclusiones. Cada sección aborda aspectos clave del estudio, proporcionando un análisis detallado y conclusiones basadas en los hallazgos.

1.1. Objetivos del estudio

El objetivo general de este estudio es analizar y evaluar la capacidad de los modelos de redes neuronales convolucionales profundas (RNCP) para detectar de manera precisa y efectiva el cáncer de mama en imágenes de mamografías. El enfoque principal será examinar su desempeño en diferentes contextos clínicos y bases de datos, asegurando la representatividad y generalización de los resultados obtenidos a poblaciones diversas. Los objetivos específicos son:

- **Comparar el desempeño de los modelos de RNCP en diferentes conjuntos de datos**: Evaluar la exactitud, precisión, sensibilidad y F1-score de los modelos entrenados con mamografías de bases de datos públicas, como el CBIS-DDSM, y de una base de datos local del Hospital General de Ensenada. Este análisis busca identificar diferencias en la detección de anomalías en diferentes entornos clínicos y tipos de población.
- **Detectar y analizar sesgos y limitaciones en los modelos**: Examinar cómo los sesgos de los datos, como el desbalance de clases, la variabilidad en la calidad de las imágenes o las características demográficas, pueden afectar el rendimiento de los modelos. Esto incluirá la identificación de patrones que limiten la generalización de los modelos en la detección de cáncer de mama en diversas poblaciones.
- **Proponer mejoras para aumentar la generalización y aplicabilidad de los modelos**: Con base en los resultados de la evaluación y análisis de los modelos, se desarrollarán recomendaciones específicas para mejorar la robustez y adaptabilidad de los modelos de RNCP. Estas sugerencias estarán enfocadas en mejorar la precisión de los modelos en la detección de cáncer de mama en diferentes poblaciones y condiciones clínicas, asegurando su eficacia en la práctica médica real.

1.2. Hipótesis

La hipótesis principal de este estudio es que los modelos de RNCP entrenados en bases de datos públicas pueden no ser igualmente precisos cuando se aplican a mastografías de una población específica como la del Hospital General de Ensenada.

2. Trabajo relacionado

2.1. Modelos de RNCP en el diagnóstico mamario

Las redes neuronales convolucionales se utilizan principalmente para la clasificación de imágenes, y su eficiencia mostrada en sus resultados, es una de las razones principales por las que el aprendizaje profundo y el aprendizaje automático han tenido un nuevo auge en la investigación. Las RNCP aprenden características discriminatorias automáticamente y su arquitectura está particularmente adaptada para aprovechar la estructura 2D de la imagen de entrada, pero lo que es más importante, una de sus características más impresionantes es que generalizan sorprendentemente bien otras tareas de reconocimiento de patrones. En los últimos años, se ha logrado un progreso significativo del reconocimiento de patrones en imágenes, en diferentes dominios [6, 7], a través de RNCP. Se ha demostrado que el desempeño de los métodos de aprendizaje profundo, en términos de precisión, para el problema de reconocimiento en imágenes puede sobrepasar el desempeño del humano [8, 9]. Sin embargo, todavía hay una serie de problemas por resolver, como la alta complejidad computacional de los algoritmos de aprendizaje (puede durar hasta semanas), el requerimiento de un conjunto grande de muestras para entrenar, el deterioro del desempeño del reconocimiento en función de los cambios en la base de datos de entrenamiento, etc. A pesar de estos retos, en 2020 se reportó un sistema basado en aprendizaje profundo, que resultó ser tan bueno como los médicos especialistas en la predicción del cáncer de mama al analizar las mastografías, en donde se explica que gracias a estas técnicas hubo una reducción en los falsos positivos y los falsos negativos [10]. Los modelos de RNCP han revolucionado el campo de la imagen médica, ofreciendo herramientas avanzadas para la detección automática de enfermedades. Diversos estudios han demostrado la capacidad de estas redes para identificar anomalías en imágenes de mamografías con una precisión comparable a la de los radiólogos expertos [6,11]. Por ejemplo, el uso de técnicas de Transferencia de Aprendizaje y Pseudocolor en un Sistema CADx para la clasificación de cáncer de mama ha mostrado resultados favorables en comparación con métodos del estado del arte, utilizando métricas de calidad como Exactitud, Especificidad, Sensibilidad y Medida-F [12].

2.2. Bases de datos de mastografías

Las bases de datos públicas, como el Digital Database for Screening Mammography (DDSM), proporcionan un recurso valioso para entrenar modelos de RNCP. Sin embargo, la variabilidad en la calidad de imagen y las características demográficas de las pacientes puede afectar la representatividad de los modelos. El Hospital General de Ensenada, por otro lado, tiene bajo su resguardo una base de datos específica de su población, lo que nos permitirá evaluar la generalización de los modelos en un entorno clínico real.

2.3. Comparación de resultados en diferentes entornos

Estudios previos han destacado las discrepancias en la eficacia de los modelos de redes neuronales convolucionales profundas (RNCP) cuando se aplican a diferentes bases de

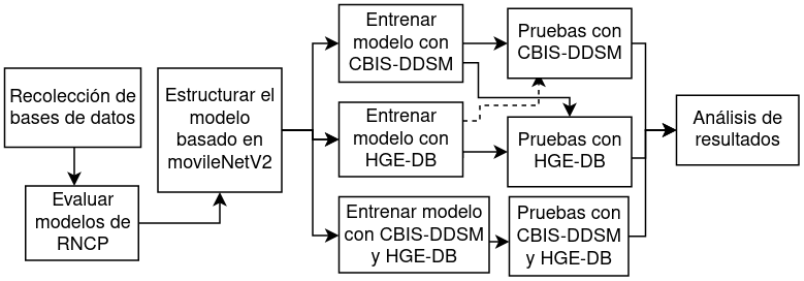

Fig. 1. Flujo del proceso metodológico utilizado en este estudio.

datos. Estas variaciones pueden deberse a diferencias en la calidad de imagen, el equipo utilizado y las características demográficas de los pacientes.

Por ejemplo, la precisión de los modelos en la detección de cáncer de mama puede variar significativamente según la calidad y diversidad demográfica de las bases de datos, con precisiones reportadas entre 74 % y 98 % [13].

En la detección de glaucoma, un modelo entrenado en un conjunto de datos específico mostró alta precisión dentro de ese conjunto pero una notable disminución cuando se aplicó a datos externos, subrayando la importancia del ajuste específico de los datos [14].

Además, en la detección y cuantificación de enfermedades pulmonares intersticiales, con enfoques de aprendizaje profundo se destaca la influencia de la preparación de datos y la integración de imágenes de múltiples canales, como la tomografía computarizada (CT), en el rendimiento de los modelos predictivos [15].

En conclusión, aunque los modelos de RNCP han demostrado ser herramientas poderosas para la detección de enfermedades a partir de imágenes médicas, su eficacia puede variar significativamente dependiendo de la base de datos utilizada y las características específicas de la población. Es crucial continuar evaluando y ajustando estos modelos para garantizar su aplicabilidad y precisión en diversos contextos clínicos.

3. Metodología

Este estudio comparativo de enfoque cuantitativo tiene como objetivo evaluar la precisión de los modelos de redes neuronales convolucionales profundas (RNCP) en distintas bases de datos de mamografías. La Fig. 1 muestra el flujo metodológico seguido para la evaluación de estos modelos. En el proceso, se entrenaron y probaron los modelos utilizando imágenes de mamografías procedentes del conjunto de datos CBIS-DDSM y del Hospital General de Ensenada. Las métricas empleadas para la evaluación incluyen exactitud, precisión, sensibilidad y el F1-score, con el fin de analizar la representatividad y eficacia de los modelos en diferentes contextos y poblaciones.

3.1. Selección de datos

En este paso se recopilan las imágenes de mastografías de 2 fuentes distintas: la base de datos pública CBIS-DDSM y la base de datos clínica del Hospital General de Ensenada (HGE-DB):

- **Base de datos CBIS-DDSM** La primera base de datos utilizada fue la CBIS-DDSM (Curated Breast Imaging Subset of DDSM), este conjunto de datos está compuesto por 6,773 imágenes de mamografías, que incluyen tanto casos benignos como malignos, con una clasificación basada en el sistema BI-RADS, es una base de datos pública que se ha utilizado extensamente en investigaciones sobre detección de cáncer de mama. En esta base de datos, las mastografías están clasificadas en 2 categorías: benignas y malignas, lo que permite entrenar y evaluar los modelos de RNCP en la detección de tumores mamarios.
- **Base de datos del hospital general de ensenada (HGE-DB)** La segunda fuente de datos es del Hospital General de Ensenada, en Baja California, México, este conjunto de datos contiene 69,950 imágenes mamográficas. Las mastografías en esta base de datos están clasificadas según la escala BI-RADS (Breast Imaging-Reporting and Data System), que evalúa el riesgo de cáncer de mama en una escala del 1 al 5. Esta clasificación proporciona una evaluación más detallada, desde normal correspondiente al BI-RADS 1 hasta altamente sospechoso de malignidad correspondiente al BI-RADS 5, ofreciendo una categorización más precisa de las imágenes mamarias.

Al combinar datos públicos y clínicos, el estudio busca comparar y evaluar la representatividad y eficacia de los modelos de RNCP en diferentes contextos y poblaciones. La selección de estas bases de datos permite analizar cómo las características de los datos afectan el rendimiento y la capacidad de generalización de los modelos.

3.2. Aprobación ética

El proyecto fue sometido y aprobado por el Comité de Ética del Hospital General de Ensenada, el cual autorizó el uso de las imágenes mamográficas del conjunto de datos HGE-DB. Garantizando de esta forma la confidencialidad de las pacientes y cumpliendo con las regulaciones éticas y de privacidad correspondientes conforme a los principios del Tratado de Helsinki [16].

3.3. Divisiones de entrenamiento y prueba

Para entrenar y evaluar los modelos, los datos fueron divididos en conjuntos de entrenamiento y prueba:

- **CBIS-DDSM**: 2,446 imágenes se utilizaron para el entrenamiento y 642 imágenes se utilizaron para la fase de prueba.

Tabla 1. Arquitectura utilizada.

Etapa	Metrica	CNN	DenseNet	Inception	MobileNet	ResNet	VGG
Entrenamiento	Exactitud	.62	.69	.66	.69	.54	.72
	Precisión	.58	.61	.62	.67	.49	.68
	Sensibilidad	.57	.82	.59	.61	.70	.72
	F1-score	.57	.70	.61	.64	.58	.70
Prueba	Exactitud	.60	.60	.63	.65	.48	.61
	Precisión	.50	.50	.55	.57	.41	.52
	Sensibilidad	.51	.74	.57	.56	.66	.55
	F1-score	.51	.60	.56	.56	.51	.53

– **HGE-DB**: Debido al desbalance de clases (con menos imágenes de las categorías de BI-RADS más altas), se seleccionaron 1,077 imágenes para el entrenamiento y 678 para la prueba.

3.4. Modelos evaluados

Durante el experimento, se probaron varios modelos de RNCP (ver Tabla 1). Entre los modelos evaluados, MobileNetV2 mostró los mejores resultados tanto en exactitud como en precisión. Otros modelos evaluados, como ResNet y VGG16, no lograron superar a MobileNetV2 en términos de rendimiento.

En el contexto médico de la detección del cáncer de mama, es fundamental que los modelos presenten altos niveles de precisión y exactitud:

– **Precisión**: Es esencial para evitar falsos positivos, es decir, la predicción de cáncer en pacientes que no lo tienen. Minimizar los falsos positivos reduce el número de pruebas adicionales invasivas y la ansiedad en las pacientes, lo que hace que la precisión sea una métrica crítica para el uso clínico.

– **Exactitud**: Aunque abarca tanto falsos positivos como negativos, una alta exactitud asegura que el modelo clasifica correctamente la mayoría de las imágenes, lo cual es crucial para su confiabilidad general en la práctica clínica. Esto asegura que tanto los casos positivos como los negativos se detecten con una alta tasa de éxito.

3.5. Descripción de los modelos de RNCP utilizados

El modelos de RNCP empleados fueron entrenados utilizando una arquitectura popular llamada mobileNet, con ajustes específicos para el aprendizaje de la misma.

MobileNetV2 es una arquitectura de red neuronal convolucional optimizada para dispositivos móviles y otros entornos con recursos limitados.

Introducida por Google en 2018 [17], mejora la eficiencia y el rendimiento de su predecesora MobileNetV1 mediante el uso de:

Tabla 2. Arquitectura utilizada.

Capa	Tipo	Salida
mobilenetv2_1.00_224	Funcional	(None, 16, 16, 1280)
Flatten	Aplanar	(None, 327680)
Fully_Connected	Secuencial	(None, 5) HGE-DB (None, 2) CBIS-DDSM

- **Bloques residuales invertidos:** Estas estructuras permiten que la información fluya más fácilmente a través de la red, mejorando la capacidad de aprendizaje y la precisión del modelo.
- **Conexiones lineales de cuello de botella:** Se utilizan para mantener la eficiencia computacional y reducir la cantidad de operaciones necesarias.

En la Tabla 2 se muestra la arquitectura de red neuronal convolucional profunda utilizada. Comienza con MobileNetV2, una red pre entrenada eficiente que extrae características de las imágenes de entrada con dimensiones de 512 × 512, produciendo una salida con dimensiones de 16 × 16 y 1280 canales de características. Luego, esta salida se aplana en un vector unidimensional de 327680 elementos mediante una capa de aplanado. Finalmente, este vector se pasa a través de una capa completamente conectada que reduce la dimensionalidad a 5 o 2 unidades, adecuada para clasificar las imágenes en una de cinco categorías posibles para el caso del clasificador BI-RADS o en 2 categorías posibles para la clasificación de las mastografías etiquetadas como benigno o maligno.

3.6. Entrenamiento del modelo

Se utilizó un modelo basado en MobileNetV2 ya pre-entrenado con ImageNet. Este modelo se entrena con la base de datos CBIS-DDSM y posteriormente se prueba utilizando un conjunto de imágenes de la misma base de datos. Esto permite evaluar la precisión y eficacia del modelo en un entorno controlado. El modelo se afina utilizando las imágenes de la base de datos HGE-DB. Este entrenamiento adicional permite al modelo adaptarse a las características específicas de la población atendida en el Hospital General de Ensenada. Se realizan pruebas del modelo entrenado con HGE-DB utilizando un conjunto de imágenes de la misma base de datos. Este paso es crucial para evaluar la representatividad del modelo en un entorno clínico real. Finalmente, se analizan los resultados obtenidos de las pruebas realizadas con ambas bases de datos.

3.7. Procedimiento de evaluación

Los modelos fueron evaluados mediante métricas estándar como exactitud, precisión, sensibilidad y F1-score. Estas métricas proporcionan una evaluación integral del rendimiento de los modelos en términos de su capacidad para identificar correctamente las anomalías en las mastografías y minimizar los errores de clasificación. La descripción de estas métricas es la siguiente:

Tabla 3. Matriz de confusión de modelo entrenado y probado con CBIS-DDSM, 2 clases.

	Clase verdadera	
Benigno	279 (73%)	102 (27%)
Malingo	115 (44%)	145 (56%)
Predición	Benigno	Maligno

Tabla 4. Métricas del modelo entrenado y probado con CBIS-DDSM, 2 clases.

Clase	**Exactitud**	**Precisión**	**Sensibilidad**	**F1-score**
Benigno	.66	.71	.73	.72
Maligno	.66	.59	.56	.57

- **Exactitud.** Mide el porcentaje de predicciones correctas realizadas por el modelo sobre el total de casos evaluados.
- **Precisión.** Proporción de verdaderos positivos entre todos los casos clasificados.
- **Sensibilidad.** Proporción de verdaderos positivos entre todos los casos positivos.
- **F1-Score.** Media armónica de precisión y sensibilidad.

Este análisis de evaluación permite determinar la eficacia y capacidad de generalización del modelo en diferentes contextos y poblaciones.

4. Resultados

Para evaluar la hipótesis de que los modelos de RNCP entrenados en bases de datos públicas pueden no ser igualmente precisos cuando se aplican a mamografías de una población específica como la del Hospital General de Ensenada, se utilizaron matrices de confusión y se realizó un análisis detallado de las métricas de rendimiento.

4.1. Desempeño de los modelos en mastografías públicas

En la Tabla 3 se presenta la matriz de confusión del modelo entrenado y probado con la base de datos CBIS-DDSM. Esta matriz muestra el número de casos correctamente e incorrectamente clasificados en las categorías benigno y maligno.

En la Tabla 4 se resumen las métricas de exactitud, precisión, sensibilidad y F1-score para cada clase, calculadas a partir de la matriz de confusión. Los resultados muestran que el modelo tiene una mayor precisión y sensibilidad en la clasificación de mastografías benignas en comparación con las malignas. La precisión y el F1-score para la clase benigna son 0.71 y 0.72, respectivamente, mientras que para la clase maligna son 0.59 y 0.57, indicando una menor capacidad del modelo para identificar correctamente las mastografías malignas. Estos hallazgos sugieren que, aunque el modelo es razonablemente efectivo en la detección de mastografías benignas, existe una necesidad de mejorar su capacidad para clasificar correctamente las mastografías malignas, posiblemente ajustando los parámetros del modelo o incorporando más datos

Tabla 5. Matriz de confusión de modelo entrenado y probado con CBIS-DDSM, 5 clases.

		Clase verdadera			
BI-RADS 1	0	0	0	0	2 (100%)
BI-RADS 2	0	63 (64%)	6 (6%)	14 (14%)	16 (16%)
BI-RADS 3	0	13 (14%)	26 (28%)	38 (41%)	16 (17%)
BI-RADS 4	0	21 (6%)	27 (8%)	216 (66%)	62 (19%)
BI-RADS 5	0	15 (12%)	14 (12%)	50 (41%)	42 (35%)
Predición	BI-RADS 1	BI-RADS 2	BI-RADS 3	BI-RADS 4	BI-RADS 5

Tabla 6. Métricas del modelo entrenado y probado con CBIS-DDSM, 5 clases.

Clase	Exactitud	Precisión	Sensibilidad	F1-score
BI-RADS 1	1	0	0	0
BI-RADS 2	.87	.56	.64	.60
BI-RADS 3	.82	.36	.28	.31
BI-RADS 4	.67	.68	.66	.67
BI-RADS 5	.73	.30	.35	.32

de entrenamiento específicos para esta clase. También se podría entrenar el modelo con las regiones de interés (ROI) contenidas en la base de datos pública, sin embargo, la base de datos proporcionada por el hospital general de ensenada no cuenta con esta misma información, por lo que, para desarrollar este análisis comparativo, ambas bases de datos deben contar con las misma información.

En la Tabla 5 se presenta la matriz de confusión del modelo entrenado y probado con la base de datos CBIS-DDSM. Esta matriz muestra el número de casos correctamente e incorrectamente clasificados con la escala BI-RADS. Se puede observar rápidamente que el modelo no aprendió a clasificar correctamente el BI-RADS 1, esto debido al desbalance drástico de esta clase.

En la Tabla 6 se resumen las métricas de exactitud, precisión, sensibilidad y F1-score para cada clase, calculadas a partir de la matriz de confusión. Los resultados muestran que el modelo tiene una mayor exactitud en la clasificación del BI-RADS 2 y 3.

Sin embargo la precisión en su clasificación es muy baja. Esto debido a que la distribución de esta base de datos fue creada para clasificar calcificaciones y masas como benignas y malignas, por lo que al tratar de redistribuir los datos ahora con la clasificación BI-RADS, esta nueva distribución de clases está desbalanceada.

4.2. Desempeño de los modelos en mastografías del Hospital General de Ensenada

En la Tabla 7 y 8 se presenta la matriz de confusión y las métricas del modelo entrenado y probado con la base de datos del Hospital General de Ensenada (HGE-DB), con las clases agrupadas según la escala BI-RADS.

Tabla 7. Matriz de confusión de modelo entrenado y probado con HGE-DB, 5 clases.

			Clase verdadera		
BI-RADS 1	70 (36%)	120 (61%)	2 (1%)	3 (2%)	1 (1%)
BI-RADS 2	55 (27%)	139 (69%)	4 (2%)	3 (1%)	0 (0%)
BI-RADS 3	44 (24%)	121 (67%)	4 (2%)	10 (6%)	2 (1%)
BI-RADS 4	8 (11%)	13 (18%)	10 (14%)	41 (56%)	1 (1%)
BI-RADS 5	0 (0%)	7 (32%)	3 (14%)	4 (18%)	8 (36%)
Predición	BI-RADS 1	BI-RADS 2	BI-RADS 3	BI-RADS 4	BI-RADS 5

Tabla 8. Métricas del modelo entrenado y probado con HGE-DB, 5 clases.

Clase	**Exactitud**	**Precisión**	**Sensibilidad**	**F1-score**
BI-RADS 1	.65	.40	.36	.38
BI-RADS 2	.52	.35	.69	.46
BI-RADS 3	.71	.17	.02	.04
BI-RADS 4	.92	.67	.56	.61
BI-RADS 5	.97	.67	.36	.47

Tabla 9. Matriz de confusión de modelo entrenado y probado con HGE-DB en grupos, 2 clases.

	Clase verdadera	
BI-RADS 1,2,3	572 (99%)	6 (1%)
BI-RADS 4,5	81 (85%)	14 (15%)
Predición	BI-RADS 1,2,3	BI-RADS 4,5

Dado el bajo rendimiento inicial, se decidió reagrupar las clases en 2 categorías: BI-RADS 1, 2, 3 (benigno) y BI-RADS 4, 5 (maligno). Esta reagrupación también se decidió realizar para permitir una comparación directa con los resultados obtenidos del modelo creado a partir de la base de datos CBIS-DDSM entrenado con 2 clases.

La Tabla 9 muestra la matriz de confusión resultante de este nuevo enfoque, observando un sobre ajuste en la primera clase, esto debido a que estas clases cuentan con un mayor número de imágenes. En la Tabla 10 se presentan las métricas de exactitud, precisión, sensibilidad y F1-score para esta nueva agrupación de clases. Los resultados muestran una mejora significativa en la precisión y sensibilidad para la categoría BI-RADS 1, 2, 3 (benigno), pero el rendimiento sigue siendo limitado para la categoría BI-RADS 4, 5 (maligno).

Los resultados indican que, aunque el modelo es altamente eficaz para clasificar las mastografías benignas (BI-RADS 1, 2, 3), tiene dificultades para identificar correctamente las mastografías malignas (BI-RADS 4, 5).

Esto subraya la importancia de considerar las características específicas de los datos de entrenamiento y su representatividad para mejorar la precisión en diferentes contextos clínicos.

Tabla 10. Métricas del modelo entrenado y probado con HGE-DB en grupos, 2 clases.

Clase	Exactitud	Precisión	Sensibilidad	F1-score
BI-RADS 1,2,3	.87	.88	.99	.93
BI-RADS 4,5	.87	.70	.15	.24

Tabla 11. Matriz de confusión de modelo entrenado con CBIS-DDSM y probado con HGE-DB, 2 clases.

	Clase verdadera	
BI-RADS 1,2,3 (Benigno)	570 (99%)	8 (1%)
BI-RADS 4,5 (Maligno)	93 (98%)	2 (2%)
Predición	Benigno	Maligno

Tabla 12. Métricas del modelo entrenado con CBIS-DDSM y probado con HGE-DB, 2 clases.

Clase	Exactitud	Precisión	Sensibilidad	F1-score
BI-RADS 1,2,3 (Benigno)	.85	.86	.99	.92
BI-RADS 4,5 (Maligno)	.85	.20	.02	.04

4.3. Comparación y discusión

Para evaluar la representatividad de los modelos de RNCP entrenados con datos públicos, se entrenó el modelo con la base de datos CBIS-DDSM y se probó con la base de datos del Hospital General de Ensenada (HGE-DB). Los resultados se presentan en las Tablas 11 y 12; En la Tabla 11 se muestra la matriz de confusión del modelo entrenado con CBIS-DDSM con 2 clases y probado con HGE-DB con 2 grupos de clases; La Tabla 12 resume las métricas de exactitud, precisión, sensibilidad y F1-score derivadas de la matriz de confusión, proporcionando una evaluación cuantitativa del rendimiento del modelo en la base de datos HGE-DB.

Los resultados indican una alta precisión y sensibilidad del modelo para las categorías benignas (BI-RADS 1, 2, 3), con un F1-score de 0.91. Sin embargo, el desempeño en la detección de categorías malignas (BI-RADS 4, 5) es notablemente bajo, con una precisión de 0.20, una sensibilidad muy baja de 0.02 y un F1-score de 0.04. Estos hallazgos sugieren que los modelos entrenados con datos de CBIS-DDSM no son suficientemente representativos cuando se aplican a datos del Hospital General de Ensenada, destacando una falta de generalización en diferentes contextos clínicos. La discrepancia en el rendimiento del modelo subraya la importancia de utilizar datos de entrenamiento que reflejen las características demográficas y clínicas específicas de la población objetivo para mejorar la precisión y eficacia de los modelos de RNCP.

En la Tabla 13 se muestra la matriz de confusión del modelo entrenado con CBIS-DDSM con 5 clases y probado con HGE-DB con 5 clases, en donde puede se puede observar de igual manera la falta de la representatividad de la clase etiquetada como BI-RADS 1. Así también, en la Tabla 14 se puede observar la falta de representatividad de las demás clases observando la baja precisión y sensibilidad en los resultados.

Tabla 13. Matriz de confusión de modelo entrenado con CBIS-DDSM y probado con HGE-DB, 5 clases.

			Clase verdadera		
BI-RADS 1	0 (0%)	126 (64%)	2 (1%)	67 (34%)	1 (1%)
BI-RADS 2	0 (0%)	133 (66%)	1 (0%)	67 (33%)	0 (0%)
BI-RADS 3	0 (0%)	103 (57%)	2 (1%)	74 (41%)	2 (1%)
BI-RADS 4	0 (0%)	38 (52%)	4 (5%)	31 (42%)	0 (0%)
BI-RADS 5	0 (0%)	8 (36%)	3 (14%)	10 (45%)	1 (5%)
Predición	BI-RADS 1	BI-RADS 2	BI-RADS 3	BI-RADS 4	BI-RADS 5

Tabla 14. Métricas del modelo entrenado con CBIS-DDSM y probado con HGE-DB, 5 clases.

Clase	**Exactitud**	**Precisión**	**Sensibilidad**	**F1-score**
BI-RADS 1	.71	0	0	0
BI-RADS 2	.49	.33	.66	.44
BI-RADS 3	.72	.17	.01	.02
BI-RADS 4	.61	.12	.42	.19
BI-RADS 5	.96	.25	.05	.08

Tabla 15. Matriz de confusión de modelo entrenado con HGE-DB y probado con CBIS-DDSM, 5 clases.

			Clase verdadera		
BI-RADS 1	0 (0%)	0 (0%)	0 (0%)	2 (100%)	0 (0%)
BI-RADS 2	0 (0%)	9 (9%)	4 (4%)	81 (82%)	5 (5%)
BI-RADS 3	1 (1%)	8 (9%)	2 (2%)	79 (85%)	3 (3%)
BI-RADS 4	1 (0%)	12 (4%)	17 (5%)	292 (90%)	4 (1%)
BI-RADS 5	1 (1%)	6 (5%)	4 (3%)	110 (91%)	0 (0%)
Predición	BI-RADS 1	BI-RADS 2	BI-RADS 3	BI-RADS 4	BI-RADS 5

Tabla 16. Métricas del modelo entrenado con HGE-DB y probado con CBIS-DDSM, 5 clases.

Clase	**Exactitud**	**Precisión**	**Sensibilidad**	**F1-score**
BI-RADS 1	.99	0	0	0
BI-RADS 2	.82	.26	.09	.13
BI-RADS 3	.82	.07	.02	.03
BI-RADS 4	.52	.52	.90	.66
BI-RADS 5	.79	0	0	0

Tabla 17. Matriz de confusión de modelo entrenado y probado con CBIS-DDSM y HGE-DB, 5 clases.

			Clase verdadera		
BI-RADS 1	82 (42%)	48 (24%)	41 (21%)	25 (13%)	0 (0%)
BI-RADS 2	67 (33%)	63 (31%)	36 (18%)	35 (17%)	0 (0%)
BI-RADS 3	59 (33%)	58 (32%)	29 (16%)	35 (19%)	0 (0%)
BI-RADS 4	6 (8%)	2 (3%)	16 (22%)	49 (67%)	0 (0%)
BI-RADS 5	2 (9%)	0 (0%)	5 (23%)	9 (41%)	6 (27%)
Predición	BI-RADS 1	BI-RADS 2	BI-RADS 3	BI-RADS 4	BI-RADS 5

En la Tabla 15 se muestra la matriz de confusión del modelo entrenado con HGE-DB con 5 clases y probado con CBIS-DDSM con 5 clases, en donde se puede observar como un posible sobre ajuste de la clase con la etiqueta BI-RADS 4, sin embargo esta clase es de las que tiene un número menor de imágenes de entrenamiento del modelo generado con HGE-DB. Así como también la representatividad de las clases debería de ser aproximadamente compatible con los resultados obtenidos anteriormente con este mismo modelo, ya que la escala BI-RADS es una escala estándar en el mundo de la radiología. Y como podemos observar tanto en la Tabla 16 como en la Tabla 8, la falta de representatividad de los datos obtenidos de distintas fuentes de datos

Para finalizar con esta sección se muestra en la Tabla 17 la matriz de confusión de los resultados obtenidos al combinar estas dos bases de datos, con la finalidad de observar el comportamiento de un modelo que es entrenado con distintas fuentes de datos, y si la representatividad del conocimiento de la misma mejora o no.

En la Tabla 18 en comparación con la Tabla 8, se puede observar una disminución en las métricas de exactitud, precisión, sensibilidad y F1-score al entrenar el modelo combinando la base de datos HGE-DB y la CBIS-DDSM, en comparación a los obtenidos al entrenar el modelo solo con HGE-DB. Esta disminución podría deberse a diversos factores:

- **Falta de representatividad de los datos**: Es posible que al combinar dos bases de datos con características poblacionales y demográficas diferentes, el modelo no sea capaz de generalizar correctamente, ya que los datos pueden estar representando diferentes distribuciones de las características clave (tipo de cáncer, calidad de imagen, etc.). Esto genera que el modelo tenga dificultades para aprender patrones comunes entre las bases de datos, afectando su rendimiento.
- **Diferencias en la calidad de las imágenes**: Las imágenes de ambas bases de datos pueden tener diferentes resoluciones, configuraciones de adquisición o protocolos médicos. Estas discrepancias en la calidad y estandarización de las imágenes pueden influir en el desempeño del modelo, que no puede adaptarse correctamente a las distintas fuentes de datos.
- **Dificultades de integración**: La combinación de bases de datos heterogéneas a veces genera "ruido" en los datos que hace más difícil que el modelo pueda identificar correctamente las anomalías. Este ruido puede surgir de diferencias en los etiquetados,

Tabla 18. Métricas del modelo entrenado y probado con CBIS-DDSM y HGE-DB, 5 clases.

Clase	Exactitud	Precisión	Sensibilidad	F1-score
BI-RADS 1	.63	.38	.42	.40
BI-RADS 2	.63	.37	.31	.34
BI-RADS 3	.63	.23	.16	.19
BI-RADS 4	.81	.32	.67	.43
BI-RADS 5	.98	1	.27	.43

los estándares de clasificación o incluso en los métodos de diagnóstico utilizados en los diferentes centros.

En conclusión, una falta de representatividad en los datos combinados es una explicación probable, aunque también puede deberse a problemas con la calidad de los datos o desbalances en las clases que no se están abordando adecuadamente.

5. Discusión y conclusiones

Los resultados de este estudio indican que los modelos de RNCP entrenados en bases de datos públicas, como CBIS-DDS, pueden no generalizar bien a poblaciones específicas, como las del Hospital General de Ensenada (HGE-DB). La comparación entre estas 2 bases de datos reveló una disminución significativa en la exactitud, precisión y sensibilidad del modelo cuando se aplicó a la base de datos HGE-DB, lo cual sugiere que las diferencias en la calidad de imagen y las características demográficas pueden contribuir a esta variabilidad en el rendimiento.

5.1. Limitaciones del estudio

Entre las limitaciones del estudio se encuentran:

- **Tamaño de la muestra:** El tamaño de la muestra de ambas bases de datos puede no ser representativo de todas las posibles variaciones en las imágenes de mastografías.
- **Diferencias en los equipos de adquisición de imagen:** Las diferencias en los mastógrafos utilizados en la adquisición de las mamografías pueden afectar la consistencia de los datos.
- **Posibles sesgos en la selección de datos:** La selección de datos para entrenamiento y prueba puede introducir sesgos que afectan la representatividad y generalización del modelo.

5.2. Implicaciones y recomendaciones

Los hallazgos resaltan la necesidad de desarrollar modelos de RNCP que consideren la diversidad poblacional y las variaciones en la calidad de las imágenes. Es crucial para la implementación efectiva de estas tecnologías en entornos clínicos diversos que

consideren estas variaciones. Para mejorar la generalización y precisión de los modelos de RNCP, se recomienda:

- **Incluir diversidad en los datos de entrenamiento:** Ampliar la base de datos utilizada para el entrenamiento del modelo para incluir una mayor diversidad de fuentes, asegurando que las características demográficas y de calidad de imagen sean representativas de las poblaciones objetivo.
- **Mitigar sesgos en el entrenamiento:** Desarrollar enfoques y técnicas para identificar y mitigar posibles sesgos en los datos de entrenamiento.
- **Estudios adicionales:** Realizar estudios adicionales que incluyan una mayor variedad de bases de datos para evaluar y mejorar la representatividad y eficacia de los modelos de RNCP en diferentes contextos clínicos.

Este estudio contribuye al entendimiento de la representatividad de los modelos de RNCP y subraya la importancia de incluir datos diversos en el entrenamiento de estos modelos. Los resultados evidencian que los modelos entrenados en bases de datos públicas mostraron una disminución significativa en precisión y sensibilidad cuando se aplicaron a mastografías del Hospital General de Ensenada, lo que reafirma la necesidad de enfoques más inclusivos en el desarrollo de estas tecnologías.

Como conclusión final, la representatividad de los modelos de RNCP es crucial para su efectividad en diferentes poblaciones. Es necesario un enfoque más diverso en el desarrollo y entrenamiento de modelos de RNCP para asegurar su generalización y precisión en entornos clínicos reales.

Referencias

1. World Health Organization.: Cáncer. (2020) https://www.who.int/news-room/factsheets/detail/cancer.
2. Instituto Nacional de Estadística y Geografía.: Estadísticas a propósito del día mundial de la lucha contra el cáncer de mama (2021) https://inegi.org.mx/app/salaDeprensa/noticia.html?id=6844.
3. Ferlay, J., Ervik, M., Lam, F.: Global Cancer Observatory: Cancer Today. International Agency for Research on Cancer (2024) https://gco.iarc.who.int/today.
4. Rodríguez-Ruiz, A., Krupinski, E., Mordang, J.J.: Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. Radiology, 290(2), 305–314 (2018) doi: 10.1148/radiol.2018181371.
5. Sawyer-Lee, R., Gimenez, F., Hoogi, A.: Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM). The Cancer Imaging Archive (2016) doi: 10.7937/K9/TCIA.2016.7O02S9CY.
6. Benítez-Mata, B., Castro, C., Castañeda, R.: Prediction of Breast Cancer Diagnosis by Blood Biomarkers Using Artificial Neural Networks. CLAIB. IFMBE, 75 (2020) doi: 10.1007/978-3-030-30648-9-7.
7. González-Lozoya, S.M., de la Calleja, J., Pellegrin, L.: Recognition of Facial Expressions based on CNN Features. Multimed Tools Applications, pp. 1–21 (2020) doi: 10.1007/s11042-020-08681-4.
8. Rajpurkar, P., Irvin, J., Ball, R.L.: Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists. PLoS Medicine, 15(11), e1002686 (2018) doi: 10.1371/journal.pmed.1002686.

9. Hannun, A.Y., Rajpurkar, P., Haghpanahi, M.: Cardiologist-level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network. Nature Medicine, 25(1), pp. 65–69 (2019) doi: 10.1038/s41591-018-0268-3.
10. McKinney, S.M., Sieniek, M., Godbole, V.: International Evaluation of an AI System for Breast Cancer Screening. Nature, 577(7788), pp. 89–94 (2020) doi: 10.1038/s41586-019-1799-6.
11. Rodríguez-Ruiz, A.: One View or Two Views: Comparison between DBT and Mammography Using an AI-Based Breast Cancer Detection System. Radiology, 290(2), pp. 493–500 (2019)
12. García-Ávila, O., Almaraz-Damián, J.A., Ponomaryov, V.: Sistema CADx para la clasificación de cáncer de mama basado en técnicas de Transfer Learning y Pseudocolor. Research in Computing Science, 150(5), pp. 65–76 (2021)
13. Thakur, A., Gupta, M., Sinha, D.K.: Transformative Breast Cancer Diagnosis Using CNNs with Optimized ReduceLROnPlateau and Early Stopping Enhancements. International Journal of Computational Intelligence Systems, 17(1), pp. 14 (2024) doi: 10.1007/s44196-023-00397-1.
14. Ko, Y.C., Chen, W.S., Chen, H.H.: Widen the Applicability of a Convolutional Neural-Network-Assisted Glaucoma Detection Algorithm of Limited Training Images Across Different Datasets. Biomedicines, 10(6), pp. 1314 (2022) doi: 10.3390/ biomedicines10061314.
15. Baba, T., Ogura, T.: Effects of Automatic Deep-Learning-Based Lung Analysis on Quantification of Interstitial Lung Disease: Correlation with Pulmonary Function Test Results and Prognosis. Diagnostics, 12(12), pp. 3038 doi: 10.3390/diagnostics12123038.
16. World Medical Association: WMA Declaration of Helsinki – Ethical principles for medical research involving human subjects. (2022) https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/
17. Sandler, M., Howard, A., Zhu, M.: Mobilenetv2: Inverted Residuals and Linear Bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)