Diagnóstico de cáncer de mama en mamografías por medio de diferenciación de lesiones utilizando radiómica y aprendizaje profundo.

Argiro Andrey Garro Espinosa, Pregrado

Universidad del Valle, 2025

Resumen

El cáncer de mama constituye la neoplasia más frecuente en mujeres mundialmente, siendo la detección temprana mediante mamografías fundamental para la supervivencia. No obstante, la interpretación visual de estas imágenes representa un proceso subjetivo que genera aproximadamente 22% de falsos positivos. Para incrementar la objetividad diagnóstica, este estudio explora la integración de métodos computacionales, fusionando características radiómicas y de aprendizaje profundo (DL) en la clasificación de lesiones mamográficas.

La metodología empleó imágenes de la base de datos CBIS-DDSM, sometidas a preprocesamiento mediante filtros Median Filter y CLAHE. Se extrajeron dos conjuntos de características: descriptores radiómicos (Pyradiomics) y características profundas (transfer learning con ResNet-50). Múltiples clasificadores SVM fueron entrenados bajo diversas configuraciones, variando preprocesamiento, tipo de lesión y optimización de hiperparámetros, utilizando validación cruzada. Finalmente, se desarrollaron modelos de ensamble (Soft Voting y Stacking) para fusionar ambos tipos de características.

Posterior a una clasificación binaria preliminar, el análisis de separabilidad mediante el método del codo y K-Means determinó entrenar clasificadores independientes para masas y calcificaciones. Los modelos finales obtuvieron AUC de 0.66 (radiómico) y 0.76 (DL) para masas, y 0.63 (radiómico) y 0.74 (DL) para calcificaciones. Los resultados en masas fueron consistentes con lo reportado por Stefano et al. (2025), mientras que los modelos de calcificaciones presentaron brechas de rendimiento del 15.22% y 9.22%, respectivamente. El hallazgo más relevante fue que el modelo de ensamble mediante Soft Voting para clasificación de masas logró una mejora del 1.5%, alcanzando AUC de 0.77, superando al modelo exclusivo de DL de Stefano et al. (2025).

En conclusión, aunque los modelos de DL demuestran mayor potencia, la radiómica permanece como fuente de información valiosa. La fusión de características radiómicas y profundas mediante técnicas de ensamble como Soft Voting permite generar descripciones más enriquecidas de las lesiones, resultando en sistemas de clasificación más robustos y precisos.

Palabras clave: Radiómica, deep learning, cáncer de mama, mamografía, machine learning.

Diagnóstico de cáncer de mama en mamografías por medio de diferenciación de lesiones utilizando radiómica y aprendizaje profundo.

Argiro Andrey Garro Espinosa

Una disertación

Enviada en cumplimiento parcial de los

Requisitos para el titulo de

Físico

en la

Universidad del Valle

2025

Pregrado de Fisica

Diagnóstico de cáncer de mama en mamografías por medio de diferenciación de lesiones utilizando radiómica y aprendizaje profundo.

Presentado por

Argiro Andrey Garro Espinosa, ,

Dirigido por

Director: PhD.C Luis Mercado Diaz de la Universidad de Connecticut

Codirector: Dr. Juan Sebastian Trujillo de la Universidad del Valle

Asesor: PhD. Hugo Posada Quintero de la Universidad de Connecticut

Universidad del Valle

2025

Agradecimientos

Agradezco a mis padres y a mi familia por su incondicional apoyo para perseguir mis sueños; a mis amigos, por creer siempre en mí, y a mis tutores, por su invaluable guía a lo largo de este arduo camino.

Tabla de contenido

Capítulo 1

Introducción

El cáncer de mama es la neoplasia más prevalente en la población femenina a nivel mundial y una de las principales causas de morbimortalidad oncológica. En Colombia, específicamente en Cali según el Registro de Población de Cáncer de Cali (RPCC) entre 2008-2012 se reportaron 23046 nuevos casos de cáncer, donde el cáncer de mama presenta una incidencia de 23.6% en las mujeres [1]. Para combatir la mortalidad de esta enfermedad, la mejor estrategia consiste en una detección temprana, es por esto que la Organización Mundial de la Salud (OMS) recomienda que las mujeres entre 50 y 69 años realicen tamizajes bianuales, siendo el método más común las mamografías, gracias a su bajo coste y disponibilidad [2]. Sin embargo, la detección oportuna se ve desafiada por la subjetividad y capacidad perceptual del observador. Esto puede llegar a resultar en errores de hasta el 22% en la histopatología dejando procedimientos invasivos innecesarios en los pacientes [2].

Adicionalmente, a este problema se suma la escasez de radiólogos especializados, según el Ministerio de Salud de Colombia, en el año 2017 se registraban cerca de 869 radiólogos en el país„ y se proyecta que esta cifra aumente a 1880 para el año 2030. Si combinamos esto con el constante crecimiento en la cobertura de salud, naturalmente la demanda de lectura e interpretación de imágenes diagnósticas también incrementa, generando así una brecha en la oferta-demanda de estos servicios especializados [3, 4].

Es por ello, que las imágenes diagnósticas, en particular la mamografía, se consolidan como herramientas esenciales para la detección temprana del cáncer de mama, ya que permiten identificar lesiones en etapas en las que el tratamiento puede ser más efectivo. No obstante, la interpretación de estas imágenes continúa enfrentando desafíos importantes: la variabilidad en la experiencia de los radiólogos, la alta carga de trabajo y la naturaleza sutil de algunas lesiones pueden afectar la precisión diagnóstica.

Ante las limitaciones en la interpretación de mamografías, la inteligencia artificial (IA) ha surgido como una tecnología de apoyo fundamental. Estudios recientes, como el proyecto alemán PRAIM, demuestran su impacto al aumentar la tasa de detección de cáncer de mama en un 17.6% sin incrementar las llamadas a revisión, lo que subraya su potencial para cerrar la brecha diagnóstica [5].

El desarrollo de estos sistemas ha evolucionado desde el uso de características diseñadas manualmente hacia modelos de IA más avanzados, impulsados por la creciente disponibilidad de datos médicos. Actualmente, la IA es clave en diversas tareas clínicas, como la segmentación de lesiones, la clasificación entre condiciones benignas o malignas y la predicción de patologías. Dos de las herramientas más prometedoras en este campo son la radiómica y el aprendizaje profundo (deep learning). La primera se basa en extraer y analizar una gran cantidad de características cuantitativas de las imágenes para alimentar modelos de aprendizaje automático. En contraste, el aprendizaje profundo utiliza redes neuronales para aprender representaciones complejas y patrones directamente de los datos, eliminando la necesidad de una selección manual de características.

La investigación en cáncer de mama en los campos de la radiómica y el *deep learning* ha crecido de manera exponencial desde 2015, como se muestra en la figura 1.1. Este auge se explica, por un lado, por la publicación del influyente artículo de revisión de Gillies et al. (2016) [6], que estableció las bases conceptuales de la radiómica, y por la aparición de herramientas como *PyRadiomics* (Van Griethuysen et al., 2017), que facilitaron su implementación. De forma paralela, el *deep learning* se consolidó con trabajos de gran impacto, entre ellos la

revisión de LeCun et al. (2015) y la propuesta de Srivastava et al. (2014), quienes introdujeron el método de *dropout*, contribuyendo a mitigar el sobreajuste y a mejorar la capacidad de generalización de las redes neuronales profundas.

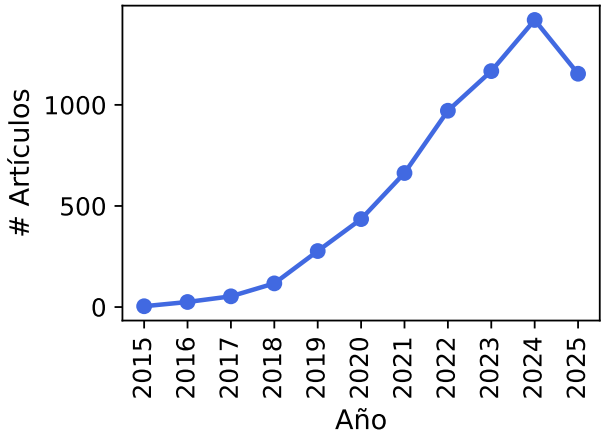


Figure 1.1: Numero de artículos por año en la base de datos Scopus, con búsquedas relacionadas con "breast cancer" AND ("radiomics" OR "deep learning") desde el 2015 hasta el 2025.

A pesar del notable auge internacional en las aplicaciones de IA para el estudio del cáncer de mama, liderado por China, país que entre 2015 y 2025 concentra casi el doble de publicaciones en comparación con Estados Unidos. Este crecimiento global contrasta marcadamente con la producción científica en países en desarrollo. En el caso de Colombia la producción científica en esta área aún es incipiente. En efecto, entre 2015 y 2025 únicamente se reportan 18 artículos publicados en revistas de Elsevier, como se muestra en la figura 1.2. Este panorama contrasta con las capacidades nacionales en investigación y desarrollo tecnológico, que permitirían una mayor participación en este campo estratégico. La baja producción científica podría estar relacionada con limitaciones en la infraestructura, la disponibilidad de bases de datos clínicas de calidad o la articulación entre grupos de investigación y el sector salud. No obstante, resulta relevante señalar que la inteligencia artificial en general se ha consolidado como un área de creciente interés a nivel global, con aplicaciones que trascienden la medicina y abarcan sectores como la industria, la educación y la economía, y cuyo desarrollo se proyecta con un crecimiento sostenido en los próximos años [7].

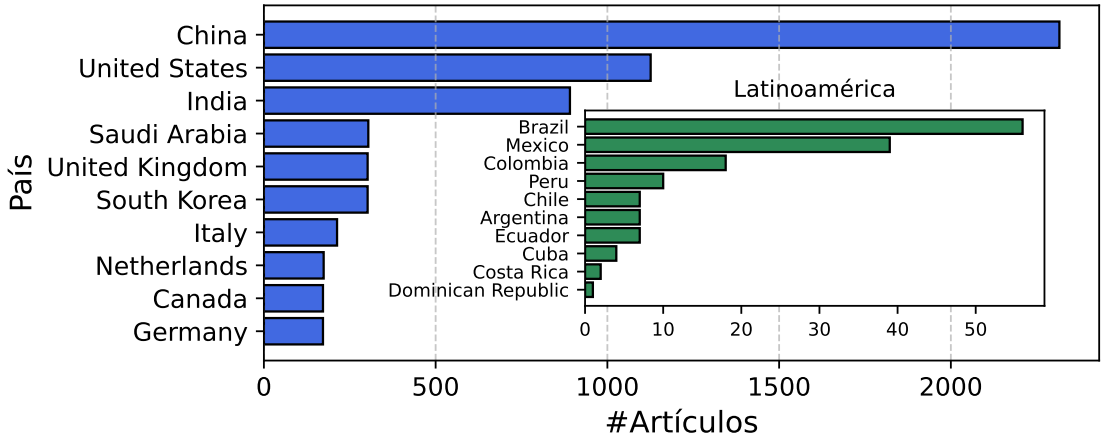


Figure 1.2: Numero de artículos publicados por país en la base de datos Scopus, con búsquedas relacionadas con "breast cancer" AND ("radiomics" OR "deep learning") desde el 2015 hasta el 2025.

Asimismo, resulta relevante analizar las afiliaciones de los artículos publicados en Colombia, como se muestra en la figura 1.3. Se observa que la Universidad Nacional y la Universidad Industrial de Santander superan en tres publicaciones a la Universidad del Valle, que únicamente ha contribuido con dos artículos en el periodo analizado. No obstante, lo más llamativo es la amplia presencia de instituciones extranjeras dentro del top 10 de mayor producción, lo cual resulta coherente si se considera que, frente al bajo volumen de publicaciones nacionales, es necesario establecer colaboraciones con instituciones que poseen mayor experiencia y trayectoria en el área para poder desarrollar investigaciones de alta calidad. En este sentido, cobra especial importancia que el presente proyecto se realice en colaboración con la Universidad de Connecticut, cuya experiencia ha sido fundamental para garantizar el rigor metodológico y la adecuada ejecución de este trabajo.

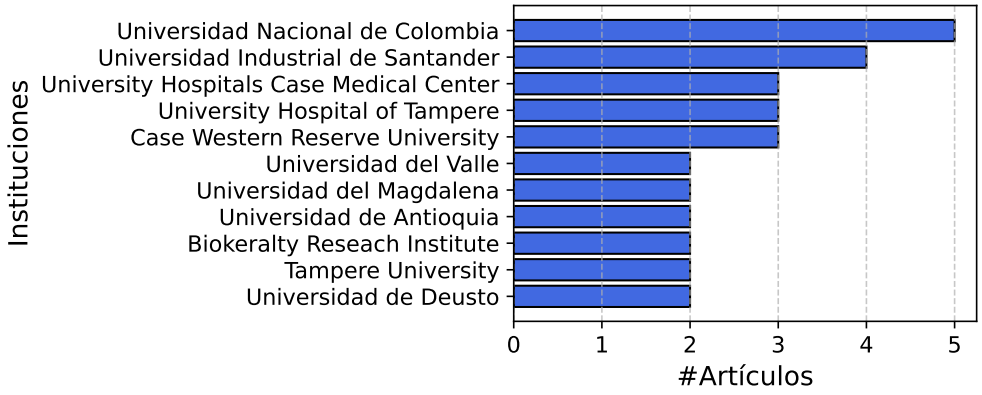


Figure 1.3: Numero de artículos por afiliación en la base de datos Scopus, con búsquedas relacionadas con "breast cancer" AND ("radiomics" OR "deep learning") desde el 2015 hasta el 2025 con respecto a Colombia.

1.0.1 Planteamiento del problema

En este contexto hemos identificado como problema central la “Falta de implementación de la IA como herramienta de apoyo que permita aumentar la eficacia en la detección temprana del cancer de mama en Colombia”. Como causas directas se encuentran:

- Bajo aprovechamiento de las capacidades nacionales para investigar IA.
- Altos índices de errores en la lectura de imágenes.
- Variabilidad en el diagnostico por la subjetividad en la interpretación de las imágenes radiológicas.
- Limitado número de estudios mamográficos para entrenar modelos de IA.

Teniendo en cuenta que la problemática se centra en encontrar alternativas que permitan incrementar la objetividad y la eficiencia en el diagnóstico temprano del cáncer de mama, en este proyecto nos proponemos dar respuesta a la siguiente pregunta de investigación: ¿Es posible clasificar lesiones malignas de benignas en mamografías usando radiómica y aprendizaje profundo , a la vez que se logran altos índices de detección que contribuyan a mejorar el diagnóstico temprano del cáncer de mama? En diversos campos del procesamiento

de imágenes médicas, la IA ha demostrado ser un mecanismo eficaz para apoyar las labores clínicas en segmentación, predicción y clasificación de patologías.

El análisis visual de imágenes médicas es esencial pero limitado por la subjetividad del observador. En este contexto, la radiómica ha surgido como una técnica avanzada para cuantificar características texturales imperceptibles al ojo humano, validando su eficacia en la clasificación del cáncer de mama mediante mamografías y RM [8, 9, 10]. Paralelamente, el aprendizaje profundo, en particular las redes neuronales convolucionales (CNN), ha demostrado un rendimiento sobresaliente en la clasificación automática de imágenes médicas [11, 12, 13]. Recientemente, los transformadores visuales han surgido como una alternativa competitiva a las CNN en el análisis de imágenes médicas, incluyendo mamografías [14, 15]. En relación con lo anterior, la combinación de enfoques radiómicos y aprendizaje profundo ha demostrado mejorar la precisión diagnóstica, integrando características radiómicas con modelos CNN preentrenados como AlexNet [16]. En este contexto, el aprendizaje por transferencia (transfer learning) se ha convertido en una estrategia clave para entrenar modelos con menos datos y menor costo computacional, lo que resulta especialmente beneficioso en países en desarrollo con recursos limitados para el entrenamiento desde cero de modelos complejos. Esta técnica permite reutilizar conocimientos adquiridos en grandes conjuntos de datos, reduciendo la necesidad de infraestructura avanzada y facilitando el desarrollo de soluciones accesibles para la detección temprana del cáncer de mama. Siguiendo esta línea, este trabajo explora arquitecturas modernas como ResNet , con el objetivo de desarrollar un modelo híbrido que optimice la clasificación de lesiones mamarias y contribuya a un diagnóstico más preciso y accesible.

1.0.2 Objetivos

Objetivo general

Desarrollar y validar un modelo computacional híbrido para la clasificación de lesiones mamarias, integrando características radiómicas y aprendizaje profundo. Se utilizarán imágenes de mamografías de bases de datos abiertas, e.g., CBIS-DDSM, y la implementación se realizará en Python, con el objetivo de mejorar la precisión en la diferenciación entre lesiones benignas y malignas.

Objetivos específicos

(1) Implementar y optimizar un modelo computacional de clasificación basado en características radiómicas extraídas de regiones de interés (ROI) en mamografías.

(2) Desarrollar un modelo computacional de clasificación basado en características profundas extraídas de un ROI en mamografías, empleando técnicas de aprendizaje profundo.

(3) Diseñar e implementar un sistema de ensamble que combine las predicciones de ambos modelos mediante técnicas de stacking, e.g., soft voting, bagging y stacking. Se evaluará la mejora en el rendimiento respecto a los modelos individuales y se validará la robustez del sistema con métricas estadísticas estándar.

1.1 Estado del arte

Diversos estudios han demostrado el potencial de la radiómica en el análisis del cáncer de mama. Padilla et al. (2022) desarrollaron un sistema de análisis parenquimal basado en recuperación de imágenes, logrando una mejora en la evaluación del riesgo con un AUC que pasó de 0,504 a 0,813. Whitney et al. (2019) mostraron que la adición de características radiómicas al tamaño de la lesión mejoró la distinción entre lesiones benignas y cánceres luminal A, elevando el AUC de 0,79 a 0,84. Li et al. (2019) combinaron características tumorales con el análisis del parénquima contralateral en mamografías, logrando también un AUC de 0,84. Por su parte, Parekh y Jacobs (2017) aplicaron un enfoque multiparamétrico con MRI y aprendizaje automático, obteniendo un AUC de 0,91, con sensibilidad del 93% y especificidad del 85% al clasificar tumores benignos y malignos. Finalmente, Li y Jiao (2022) desarrollaron un modelo supervisado de clasificación mamográfica basado en radiómica, evidenciando mejoras en la detección del cáncer de mama.[8, 17, 18, 10, 19].

En base a lo anterior, las redes neuronales convolucionales han sido utilizadas en varios estudios de cáncer de mama. Abunasser et al. aplicaron el algoritmo Xception logrando excelentes resultados con precisión y recall del 97.60% en la clasificación de ocho tipos de tumores mamarios[11]. En un estudio posterior, los mismos autores desarrollaron un modelo BCCNN que alcanzó un F1-score de 98.28%, superando a arquitecturas preentrenadas como ResNet50 (98.14%) y VGG16 (97.67%), especialmente con imágenes de magnificación 400X [12]. Lu et al. exploraron un enfoque multimodal combinando radiómica y aprendizaje profundo con imágenes de ultrasonografía, mamografía y resonancia magnética, logrando una precisión de 0.968 mediante estrategias de fusión de características [20]. Por su parte, Falconi et al. investigaron técnicas de transfer learning y fine tuning, demostrando que VGG16 con fine tuning alcanzó el mejor rendimiento con un AUC de 0.844 en la base de datos CBIS-DDSM [21].

Del análisis de la literatura revisada se desprende que los modelos de aprendizaje profundo tienden a ofrecer un rendimiento superior en la clasificación del cáncer de mama, en

comparación con los enfoques basados exclusivamente en radiómica.

No obstante, la investigación más reciente apunta hacia una tercera vía: los **modelos híbridos** que buscan integrar ambas metodologías. La premisa de estos enfoques es que la combinación de las características profundas, extraídas automáticamente, con las características radiómicas, explícitas y cuantificables, crea una representación más completa de la lesión, lo que puede conducir a un mejor desempeño predictivo. El potencial de esta sinergia se refleja en diversos trabajos, como se resume en la **Tabla 1.1**.

Autores	**Año**	**Métodos**	**Tipos de Imagen**	**Resultados principales**
Guoxiu Lu et al. [20]	2024	Radiomica, Deep learning, (ensemble stacking)	US, FFMD, MRI	AUC hasta 0.997; especificidad y sensibilidad del 100% con estrategias de fusión tardía.
Meredith A. Jones et al. [22]	2023	Radiomics, aprendizaje profundo (DTL), fusión de vistas bilaterales e ipsilaterales	FFMD (vistas CC y MLO)	AUC de 0.876 con fusión multivista; mejor que métodos basados en una sola vista.
Tariq Mahmood et al. [23]	2024	Grad-CAM, radiomics y aprendizaje profundo	Mamografía multiparamétrica	Se destaca el uso de Grad-CAM para interpretabilidad.
Volkan M. Tiryaki et al. [24]	2024	Ensemble de aprendizaje profundo (ResNet50, NASNet-Large, Xception), selección genética de características	Mamografía (BCDR)	AUC de 0.9089; mejor rendimiento en clasificación de masas en BCDR hasta la fecha.
Alessandro Carriero et al. [25]	2024	Revisión narrativa de avances en DL para cáncer de mama	FFMD, US, MRI	Identificación de desafíos clave para la adopción clínica de IA en imágenes médicas.
Xinyu Zhang et al. [26]	2021	Radiomica y características profundas integradas	FFMD	AUC de 94.6%, precisión de 96.4%; combinación mejora notablemente la clasificación.
Ahmet H. Yurttakal et al. [27]	2021	Ensemble de boosting y DL, radiomics refinados	DCE-MRI	Precisión de 94.87%; AUC de 0.9728 con recall perfecto (1.0).
Yanhua Cui et al. [28]	2021	Fusión de características clínicas, radiomics y DL (VGG16, Inception-V3)	FFMD	AUC de 0.961 en cohortes de prueba, mejorando el modelo clínico tradicional.

Table 1.1: Resumen de artículos sobre métodos de fusion de radiómica y DL y modalidades en imágenes médicas para cáncer de mama. Abreviaturas: mamografía digital (FFMD), imagen de resonancia magnética (MRI), ultrasonido (US).

Capítulo 2

Fundamento teórico

2.1 Radiómica

La radiómica representa un paradigma emergente que transforma imágenes médicas en datos cuantitativos de alta dimensionalidad, permitiendo caracterizar la fisiopatología subyacente mediante análisis computacional avanzado[6] . Este enfoque metodológico se fundamenta en la hipótesis de que las características extraídas de imágenes médicas reflejan procesos biológicos y pueden correlacionarse con resultados clínicos específicos[29].

Un ejemplo de esta correlación fue evidenciado por Ogbonnaya et al. (2023), quienes encontraron que ciertas características texturales extraídas de imágenes de resonancia magnética ponderadas por perfusión (bpMRI), como el histograma y la matriz de co-ocurrencia de niveles de gris (GLCM), se correlacionaban con la expresión de genes relacionados con hipoxia (ANGPTL4, VEGFA y P4HA1) en cáncer de próstata localizado. Además, un análisis radiogenómico posterior demostró que la sobreexpresión de estos genes se asociaba con una menor sobrevivida global, con una mediana de 81.11 meses en pacientes con alteraciones frente a 133.00 meses en aquellos sin alteraciones. Esta capacidad de inferir información biológica a partir de imágenes convierte a la radiómica en una herramienta clave para fortalecer el vínculo entre diagnóstico por imagen y biología molecular, facilitando así el de-

sarrollo de una medicina más personalizada, predictiva y basada en evidencia cuantitativa [30].

El análisis radiómico sigue un flujo de trabajo sistemático que comprende varias etapas fundamentales para garantizar la calidad y reproducibilidad de los resultados. En este proyecto se busca realizar la clasificación de lesiones mamarias, un esquema general del flujo radiómico sería como en la siguiente figura.[31].

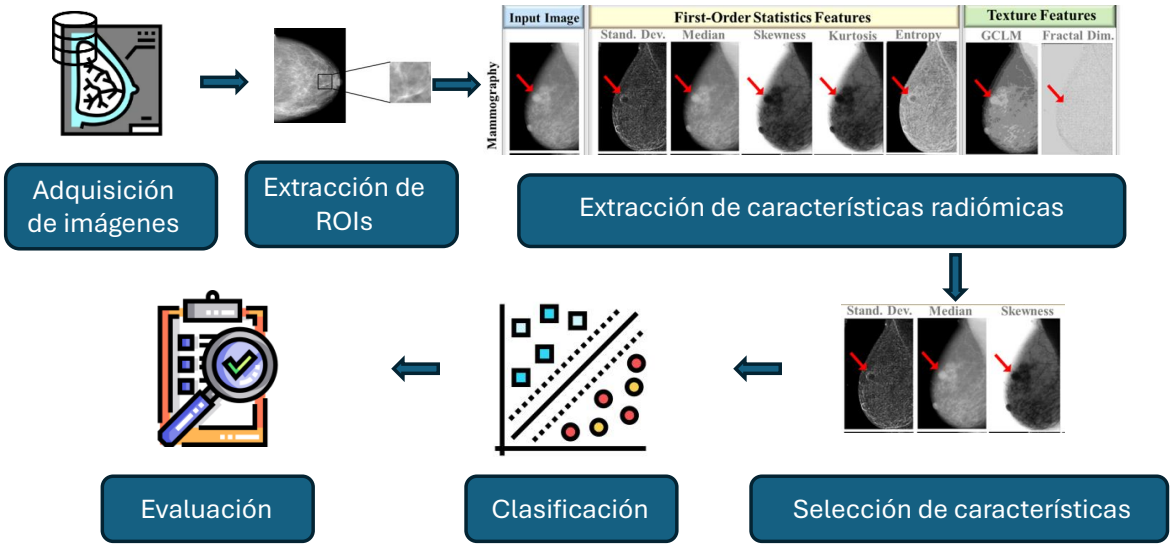


Figure 2.1: Flujo de trabajo en radiómica. 1) La adquisición de las imágenes del estudio, 2) La extracción de regiones de interés (ROI). 3) La extracción de las características radiómicas, descriptores matemáticos de las propiedades de las imágenes. 4) Selección de características. 4) Entrenamiento de un modelo de clasificación

Aunque el flujo de trabajo en radiómica pueda parecer robusto a primera vista, la disciplina todavía enfrenta desafíos importantes que deben resolverse para que alcance su verdadero potencial como herramienta de apoyo clínico. En la siguiente sección analizamos con mayor detalle los problemas de reproducibilidad y estandarización en radiómica.

2.1.1 El desafío de la reproducibilidad

De acuerdo con Demircioglu et al.[32], la radiómica enfrenta un desafío central: la reproducibilidad. Este concepto implica que resultados similares deberían obtenerse bajo condiciones equivalentes, como escanear dos veces al mismo paciente en un intervalo corto. Cuando los hallazgos no son reproducibles, los estudios resultan erráticos, poco confiables y clínicamente inútiles, pues generan falsos positivos que imposibilitan su aplicación en la práctica médica. En lo que sigue, se resume y discute lo planteado por dichos autores sobre la problemática de la reproducibilidad en radiómica.

La reproducibilidad en radiómica puede dividirse en dos grandes áreas: reproducibilidad en la adquisición de imágenes y reproducibilidad estadística. La primera se refiere al proceso de obtención de imágenes y extracción de características, mientras que la segunda está relacionada con la construcción de modelos predictivos mediante aprendizaje automático. Si la adquisición de imágenes no es reproducible, ninguna técnica estadística posterior puede compensarlo, reflejando el principio de "garbage in, garbage out".

La reproducibilidad de las imágenes en radiómica depende de múltiples factores. Los parámetros de adquisición, como el tamaño del vóxel y las técnicas de reconstrucción, junto con las diferencias entre equipos y fabricantes, afectan de manera significativa la estabilidad de las características. Incluso pequeños cambios pueden alterar los resultados de forma considerable. Aunque las técnicas de armonización ayudan a reducir estas variaciones, su efecto es limitado. Además, la segmentación de las regiones de interés introduce variabilidad intra- e interobservador, lo que impacta directamente la consistencia de los descriptores extraídos. Por estas razones, las características radiómicas no son completamente objetivas, ya que aún persiste un grado de subjetividad, sobre todo en la delimitación de la región de interés.

Incluso las definiciones de características presentan inconsistencias, dificultando la estandarización. Rasgos aparentemente simples como la esfericidad pueden diferir según la fórmula utilizada para calcularlos, lo que motivó la creación de la Image Biomarker Standard-

isation Initiative (IBSI). Sin embargo, no todos los programas cumplen con estos estándares, y aún los descriptores estandarizados pueden mostrar discrepancias. Asimismo, la aplicación de filtros de preprocesamiento añade otra fuente de variabilidad, ya que no está claro qué filtros son los más adecuados o si realmente contribuyen a mejorar la capacidad predictiva.

En el plano de la reproducibilidad estadística, los datos radiómicos suelen ser de alta dimensión (más características que muestras) y altamente correlacionados, lo que incrementa el riesgo de identificar patrones espurios y producir falsos positivos. Para mitigar esto se aplican técnicas de selección de características, que buscan reducir la dimensionalidad reteniendo solo los descriptores relevantes. Existen métodos simples (p. ej., correlación de Spearman, t-test) que son más robustos y reproducibles, y métodos complejos (p. ej., LASSO, mRMR, Boruta) que pueden capturar dependencias entre variables, aunque con mayor costo computacional y, en muchos casos, sin mejoras sustanciales en el desempeño.

Finalmente, ni la selección de características ni los clasificadores ofrecen una solución definitiva. Los métodos de selección son inherentemente inestables, y la exclusión o inclusión de pocos casos puede modificar sustancialmente los resultados. Por su parte, los clasificadores enfrentan directamente los problemas de alta dimensionalidad y dependen de supuestos y de hiperparámetros cuyo ajuste exhaustivo rara vez es posible debido a las limitaciones de recursos computacionales. Esto implica que los modelos obtenidos pueden estar lejos de ser los óptimos, afectando la reproducibilidad y la aplicabilidad clínica de la radiómica.

A esto se suma que la estandarización de las imágenes es poco práctica y la variabilidad estadística es inherente a los conjuntos de datos de alta dimensión, lo que dificulta aún más la integración clínica. En este contexto, resulta indispensable avanzar hacia prácticas más rigurosas de intercambio de datos y código, así como hacia el desarrollo de bases de datos grandes y representativas. Una de ellas es *CBIS-DDSM*, que constituye un punto de partida relevante para comprender cómo estas consideraciones impactan en la construcción de un flujo de trabajo radiómico.

2.1.2 Adquisición y preprocesamiento de las imágenes.

Base de datos CBIS-DDSM

El *Curated Breast Image Subset of Digital Database for Screening Mammography* (CBIS-DDSM) constituye un conjunto de datos curado específicamente diseñado para resolver los desafíos de estandarización y reproducibilidad que enfrentan los sistemas de diagnóstico asistido por computadora (CADx) en mamografía.

Esta base de datos conserva las anotaciones originales del DDSM, enriquecidas con correcciones y validaciones realizadas por especialistas en mamografía. La distribución de las imágenes en conjuntos de entrenamiento y evaluación se realizó mediante una partición 80/20, preservando la distribución de dificultad diagnóstica según la clasificación BI-RADS en ambos subconjuntos. La segmentación de las ROIs fue realizada mediante algoritmos automáticos de segmentación. Esta y más información sobre el CBIS-DDSM puede ser consultada en Lee et al (2017) [13].

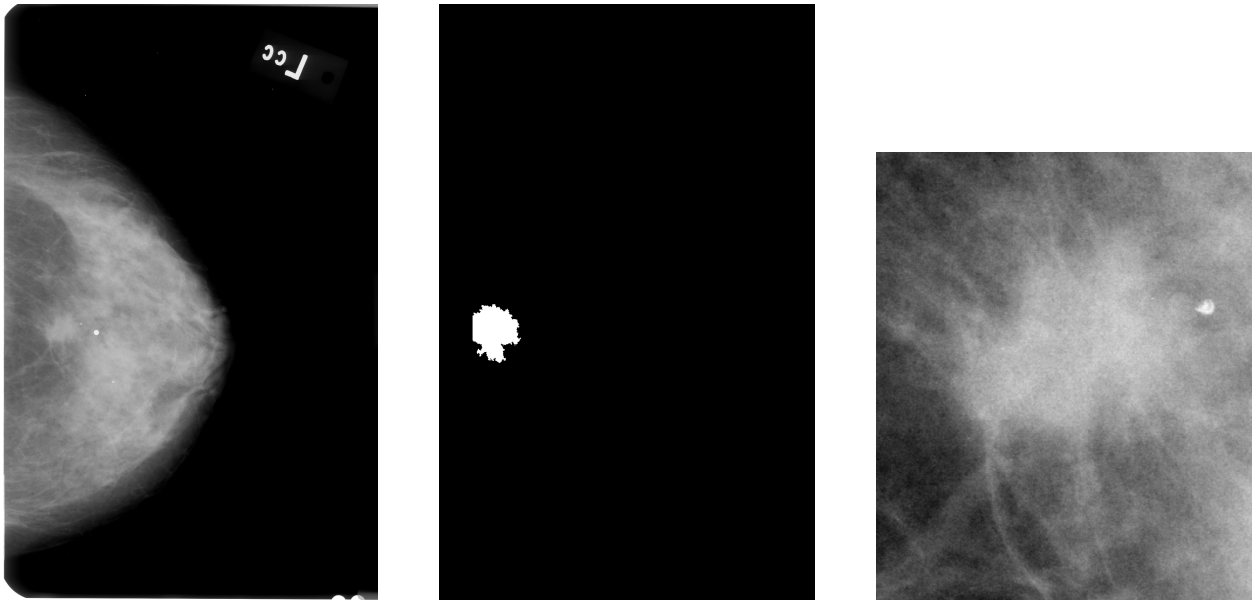
La principal ventaja de la base de datos CBIS-DDSM para la clasificación de lesiones mamarias mediante radiómica radica en la inclusión de Regiones de Interés (ROI) previamente definidas. Como se discutió en la sección anterior, la segmentación de estas regiones no solo constituye un paso fundamental en el flujo de trabajo radiómico, sino que también es un elemento clave para garantizar la reproducibilidad de los resultados del estudio.

Adicionalmente, un aspecto invaluable de esta base de datos es la disponibilidad de etiquetas patológicamente verificadas para cada lesión. Esto permite el desarrollo y entrenamiento de modelos de aprendizaje supervisado, facilitando la creación de clasificadores basados en radiómica y en deep learning.

	Benignas	**Malignas**
Entrenamiento de calcificaciones	552	304
Prueba de calcificaciones	112	77
Entrenamiento de masas	387	361
Prueba de masas	135	87

Table 2.1: Número de anormalidades en los conjuntos de entrenamiento y prueba en la base de datos CBIS-DDSM.

Las imágenes mamográficas de las anomalías del CBIS-DDSM se proporcionan con tres componentes principales: la imagen mamográfica original, las máscaras correspondientes a las regiones de interés (ROI) y una imagen recortada que consiste en un rectángulo que circunscribe la ROI. Esta última suele ser suficiente para el entrenamiento de sistemas CADx.



(a) Imagen completa **(b)** Mascara de la lesión **(c)** Imagen recortada

Figure 2.2: Descripción general del conjunto de imágenes de la base CBIS-DDSM. La imagen mamográfica completa (a) un ROI que enmarca la lesión a clasificar (b) y la imagen recortada producto de aplicar la mascara a la imagen original (c).

A pesar de que CBIS-DDSM es una versión mejorada de DDSM, aún conserva limitaciones significativas. Aunque utiliza imágenes digitalizadas en formato DICOM y segmentaciones mejoradas, sigue derivando de mamografías en película con escaneos de aproximadamente 42–50 µm, lo que afecta la calidad y estandarización de las imágenes[33]. Además, varios estudios reportan inconsistencias como nombres de archivo que no coinciden con los registros en los CSV asociados, imágenes volteadas horizontalmente y discrepancias en la resolución entre imágenes y sus máscaras, lo que complica su procesamiento y alineación[34]. También se detecta desbalance en la distribución de clases, lo cual podría sesgar los modelos hacia la clase mayoritaria si no se aplica una corrección adecuada[35]. Finalmente, se han observado artefactos visuales y elementos redundantes en algunas imágenes, producto de la tecnología

de mamografía por película, que introducen ruido no clínico y pueden afectar la extracción de rasgos precisos[34].

Preprocesamiento

El preprocesamiento de imágenes constituye una etapa fundamental en las tareas de clasificación automática, ya que permite reducir el ruido inherente a las imágenes mamográficas y mejorar su calidad para el posterior análisis computacional. Diversos autores han demostrado que la implementación de protocolos de preprocesamiento adecuados puede incrementar significativamente la precisión de los algoritmos de clasificación [36].

En el contexto de este trabajo, las imágenes del conjunto CBIS-DDSM serán sometidas a un protocolo de control basado en la normalización de intensidades.

Adicionalmente, se implementará un preprocesamiento mediante la combinación de algoritmos de median filter (MF) y contrast limited adaptive histogram equalization (CLAHE). Esta estrategia combinada ha sido validada en múltiples estudios para la clasificación de tejidos malignos y benignos en mamografía, mostrando mejoras consistentes en la diferenciación de patrones patológicos [36].

La aplicación del median filter permite la reducción del ruido impulsivo presente en las imágenes digitales, mientras que el algoritmo CLAHE mejora el contraste local adaptándose a las características específicas de cada región de la imagen. Esta combinación resulta particularmente efectiva en imágenes mamográficas, donde las diferencias sutiles en densidad tisular pueden ser determinantes para la clasificación correcta.

2.1.3 Segmentación de los regiones de interés (ROI)

La segmentación de las Regiones de Interés (ROI) constituye un pilar fundamental en el análisis radiómico. La delimitación precisa de estas áreas es esencial, ya que define el espacio a partir del cual se extraerán las características cuantitativas.

Tradicionalmente, esta segmentación se realizaba de forma manual por radiólogos expertos.

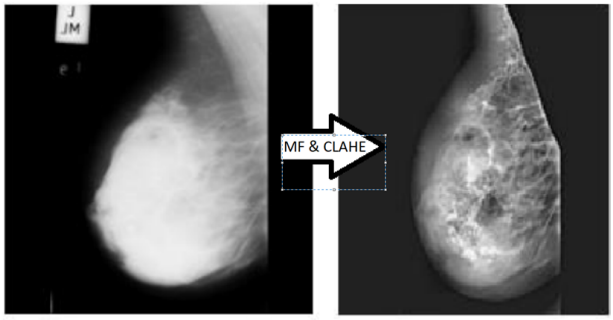


Figure 2.3: A la izquierda la imagen mamográfica original, a la derecha la imagen mamográfica después de remover las etiquetas, el musculo pectoral y aplicar los filtros MF y CLAHE [36]

Sin embargo, esta práctica conllevaba un alto grado de variabilidad interobservador, lo que afectaba la reproducibilidad de los resultados. En la actualidad, la adopción de métodos automáticos y semiautomáticos ha transformado significativamente este proceso. Estas técnicas modernas combinan enfoques como la umbralización adaptativa, modelos basados en contornos activos (por ejemplo, "snakes") y algoritmos de crecimiento de regiones.

Específicamente en el contexto de las imágenes mamarias, se han implementado exitosamente enfoques avanzados que integran Redes Neuronales Convolucionales (CNN). Estas CNN son entrenadas para reconocer bordes y estructuras morfológicas características de lesiones sospechosas. La implementación de estos métodos avanzados no solo mejora drásticamente la precisión de la segmentación, sino que también incrementa su reproducibilidad y eficiencia, aspectos vitales para su aplicación en entornos clínicos y estudios multicéntricos.

En síntesis, una segmentación precisa es indispensable porque permite capturar de manera más fiel la heterogeneidad tumoral y otros patrones complejos inherentes a la ROI. Esta fidelidad en la delimitación contribuye de forma significativa a la calidad general y a la confiabilidad del análisis radiómico, sentando las bases para la derivación de conclusiones válidas.

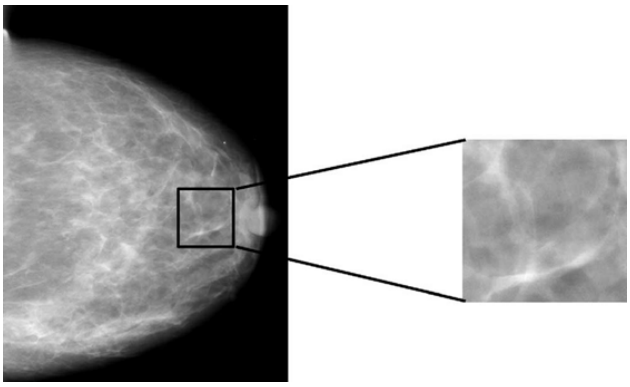


Figure 2.4: Un ROI de muestra seleccionado de la región central de la mama detrás del pezón en una mamografía digitalizada.[37]

2.1.4 Extracción de características radiómicas

Este proceso consiste en la derivación de mediciones cuantitativas objetivas a partir de las ROI segmentadas. Estas mediciones, conocidas como características radiómicas, se distinguen fundamentalmente de las características "semánticas" tradicionales, las cuales dependen inherentemente de la interpretación subjetiva de los radiólogos y se basan en descripciones cualitativas o semicuantitativas. Por ejemplo, un radiólogo puede describir una lesión mamaria como "irregular con márgenes espiculados", un nódulo pulmonar como "redondeado y homogéneo", o el patrón de realce de un tumor como "intenso y heterogéneo". Si bien estas descripciones son esenciales en la práctica clínica diaria, su naturaleza visual y dependiente del juicio humano introduce un grado de variabilidad interobservador (entre diferentes radiólogos) e intraobservador (en observaciones repetidas por el mismo radiólogo). En contraste, las características radiómicas son cuantitativas, lo que les brinda un cierto grado de objetividad, pero siguen presentando desafíos que permitan un análisis reproducible y estandarizado.

Esta objetividad radica en que las características radiómicas son descriptores estadísticos construidos manualmente, a diferencia de los modelos de deep learning, donde las representaciones se aprenden automáticamente durante el entrenamiento. En otras palabras, mientras que las redes profundas descubren patrones de manera implícita, la radiómica se apoya en atributos cuantificables y explícitos que capturan propiedades de las imágenes.

Una forma sencilla de entender este enfoque es pensar en cómo describimos una fotografía común: podemos hablar de su brillo general, de los contrastes entre zonas claras y oscuras, de la textura que transmite una superficie rugosa o lisa, o de la forma de los objetos que aparecen en ella. De manera análoga, la radiómica busca traducir las imágenes médicas en números que reflejan características de intensidad, forma, textura y patrones más complejos.

Con este marco conceptual, puede entenderse mejor que las características radiómicas se agrupan en cuatro categorías principales, cada una diseñada para aportar información complementaria sobre la heterogeneidad tumoral.

- Características de forma: Describen la morfología tridimensional del tumor, incluyendo medidas como el volumen, la esfericidad o la compacidad, que pueden reflejar aspectos de su crecimiento y agresividad de manera cuantificable, a diferencia de las descripciones semánticas como "redondeada" o "lobulada".

- Estadísticas de primer orden (histograma): Analizan la distribución de intensidades de los píxeles dentro de la ROI sin considerar su disposición espacial. Incluyen métricas como la media, la desviación estándar, la asimetría y la curtosis, que resumen propiedades globales de la distribución de señales, cuantificando la "homogeneidad" o "heterogeneidad" semántica.

- Estadísticas de segundo orden (textura): Evalúan las relaciones espaciales entre pares de píxeles, revelando patrones de heterogeneidad local que son difíciles de apreciar visualmente. Ejemplos comunes incluyen la matriz de co-ocurrencia de niveles de gris (GLCM), la matriz de tamaño de zona de nivel de gris (GLSZM) y la matriz de dependencia de nivel de gris (GLDM), que permiten cuantificar la "textura gruesa" o "fina" observada subjetivamente.

- Estadísticas de orden superior: Capturan patrones espaciales más complejos y repetitivos, y suelen derivarse de la aplicación de filtros o transformaciones matemáticas

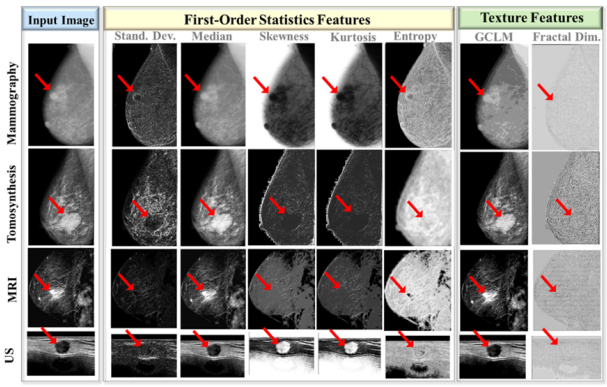


Figure 2.5: Características radiómicas voxelizadas. Se muestran ejemplos de características radiómicas extraídas de imágenes radiológicas con lesiones malignas (señaladas con flechas rojas). En este caso, se calculan tanto características de primer orden (como desviación estándar, mediana, asimetría, curtosis y entropía) como características de textura (matriz de co-ocurrencia de niveles de gris – GLCM y dimensiones fractales), a partir de imágenes obtenidas por mamografía, tomosíntesis, resonancia magnética y ecografía [39].

sobre la imagen original, extrayendo información que va más allá de la percepción visual directa.

En la sección características de textura, se detallan mas a fondo las características de segundo orden. Para mas información sobre las definiciones matemáticas de las características radiómicas se puede consultar el *Image biomarker standardisation initiative* [38].

La extracción de estas características permite cuantificar la heterogeneidad tumoral y otros patrones relevantes que pueden no ser evidentes a simple vista o ser ambiguos en la descripción semántica, proporcionando una base robusta para el desarrollo de modelos predictivos y pronósticos más precisos y generalizables. [31]. Estas características normalmente dan un valor numérico tras ser calculadas como un promedio sobre el ROI, pero también es posible extraer las características a nivel voxel,lo cual permite obtener una visualización de las mismas como se puede ver en la figura 2.5.

2.1.5 Selección de características

Una vez extraídas las características radiómicas, el siguiente paso es la selección de características. Esta etapa consiste en identificar los biomarcadores más relevantes entre la vasta

cantidad producida durante la extracción. Para ello, se distinguen tres métodos principales [40]:

- Filtro: El método filtro se basa en las características intrínsecas de los datos analizados y evalúa las características de manera independiente a cualquier algoritmo de aprendizaje. Este método utiliza matrices métricas obtenidas directamente de los datos, eliminando la necesidad de retroalimentación por parte de los algoritmos de aprendizaje para ajustar la selección de características. La selección se determina según la dispersión o relevancia de las características, y las no importantes se filtran utilizando un umbral o un número predeterminado. La creciente disponibilidad de datos masivos de alta dimensión ha incrementado el interés en los modelos filtro, los cuales no están sesgados hacia un modelo de aprendizaje específico, por lo que su estructura siempre resulta simple.

- El método wrapper se distingue del método filtro porque utiliza un algoritmo de aprendizaje predeterminado para evaluar e identificar un subconjunto de características. En lugar de basarse en métricas independientes, el método wrapper considera el algoritmo de selección de características como parte del proceso de aprendizaje y utiliza el rendimiento de clasificación como criterio de evaluación de la importancia de las características. Evaluar un subconjunto de características con un algoritmo de aprendizaje automático permite detectar interacciones entre ellas, lo que conduce a la selección de un subconjunto óptimo de características para el modelo. El método wrapper involucra tanto la búsqueda de subconjuntos de características como las métricas de evaluación. La primera se encarga de generar subconjuntos candidatos de características, mientras que la segunda entrena y evalúa un modelo con cada subconjunto utilizando un conjunto de validación para identificar el óptimo.

- Los métodos embedded combinan los beneficios de los métodos filtro y wrapper al integrar la selección de características dentro del propio proceso de construcción del

modelo. Estos métodos aprovechan de manera efectiva las ventajas de ambos enfoques: (1) son significativamente menos costosos computacionalmente que los métodos wrapper, ya que no requieren múltiples ejecuciones del modelo de aprendizaje para evaluar las características; y (2) interactúan directamente con el modelo de aprendizaje. La principal diferencia entre los métodos wrapper y embedded radica en el proceso de utilización de las características candidatas. El método wrapper entrena inicialmente el modelo de aprendizaje utilizando las características candidatas y posteriormente evalúa las características a través de la selección de estas con el modelo. En contraste, el método embedded selecciona las características durante la construcción del modelo, sin necesidad de una evaluación adicional mediante un proceso separado de selección de características.

Las ventajas y desventajas de cada método se pueden ver en la siguiente tabla 2.2

Coeficiente de Correlación de Pearson

En estadística, el coeficiente de correlación de Pearson es una medida de dependencia lineal diseñada para cuantificar el grado de relación entre dos variables, siempre que ambas sean cuantitativas y continuas. Se define como un índice adimensional cuyo valor se encuentra siempre en un rango de $[-1, 1]$, lo que facilita su interpretación al ser independiente de la escala de medida de las variables.

La fórmula para el coeficiente de correlación poblacional (ρ) se define como la covarianza entre las dos variables dividida por el producto de sus desviaciones estándar:

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Donde σ_{XY} es la covarianza de las variables (X, Y), y σ_X y σ_Y son las desviaciones estándar de X y Y respectivamente.

La interpretación del coeficiente permite entender tanto la fuerza como la dirección de la relación lineal entre las variables:

Categoría	Ventajas	Desventajas
Filter	• Cálculo más eficiente • Evitan de manera efectiva el sobreajuste • Independientes de los algoritmos de aprendizaje	• Ignoran la interacción con los algoritmos de aprendizaje • Debilitan la capacidad de ajuste del modelo
Wrapper	• Simples de implementar • Interactúan con los algoritmos de aprendizaje • Modelan dependencias entre características	• Riesgo de sobreajuste • La selección depende del algoritmo de aprendizaje • Requieren un gran número de cálculos
Embedded	• Interactúan con los algoritmos de aprendizaje • Menor complejidad que los métodos Wrapper • Cálculo más eficiente	• La selección depende del algoritmo de aprendizaje • Aumentan la carga durante el entrenamiento del modelo

Table 2.2: Ventajas y desventajas de los métodos Filter, Wrapper y Embedded.

- Un valor de **+1** indica una correlación positiva perfecta. A medida que una variable aumenta, la otra lo hace en una proporción constante.

- Un valor de **-1** representa una correlación negativa perfecta. Cuando una variable aumenta, la otra disminuye en una proporción constante.

- Un valor cercano a **0** sugiere la ausencia de una relación *lineal*. Es fundamental destacar que esto no implica que las variables sean independientes, ya que podrían existir otro tipo de relaciones no lineales entre ellas.

Clasificación de Características Radiómicas y Modelado Predictivo

Este proceso final en el flujo de trabajo radiómico tiene como objetivo cuantificar el tejido de interés y generar predicciones clínicas relevantes. Partiendo del conjunto de características previamente extraídas, cada imagen mamográfica es representada como un vector numérico en un espacio hiperdimensional. Esta representación estructurada es crucial, ya que facilita la aplicación de algoritmos de aprendizaje automático (*Machine Learning*), donde cada vector corresponde a una muestra individual y sus dimensiones reflejan propiedades cuantificables del tejido, tales como textura, forma, intensidad o patrones estadísticos.

En el presente trabajo, nos enfocamos exclusivamente en el uso de **clasificadores supervisados**. Esta elección se fundamenta en la disponibilidad de datos previamente etiquetados, es decir, imágenes mamográficas asociadas a diagnósticos clínicamente establecidos. Los clasificadores supervisados requieren un conjunto de entrenamiento que contenga ejemplos con clases conocidas; su objetivo es aprender una función de decisión o patrón que, una vez establecido, permita categorizar nuevos casos de forma automática con alta precisión.

Entre los métodos de clasificación supervisada más ampliamente utilizados en radiómica, y que se consideran en este estudio, se destacan las máquinas de vectores de soporte.

Máquinas de vectores de soporte (SVM)

La máquinas de vectores de soporte (SVM) es un modelo lineal para problemas de clasificación y regresión [41]. Es ampliamente aplicada en el campo de las predicciones médicas. Un clasificador SVM realiza una clasificación binaria, es decir, separa un conjunto de vectores de entrenamiento mapeando las dos clases diferentes $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$, donde $x_i \in \mathbb{R}^d$ denota los vectores en un espacio de características de dimensión d y $y_i \in \{-1, +1\}$ es una etiqueta de clase. Para este estudio, se estudió el modelo de máquina de soporte vectorial con función de base radial (RBF) para la clasificación de conjuntos de datos de cáncer de mama. La noción matematica de la máquina de soporte vectorial es separar los vectores de entrenamiento para las dos clases diferentes. Las siguientes ecuaciones utilizaron

la forma general para el clasificador de máquina de soporte vectorial.

$$\sum_{i=1}^{n} \infty_i - \frac{1}{2} \sum_{i,j=1}^{n} \infty_i \infty_j y_i y_j K(x_i, x_j) \tag{2.1}$$

$$\sum_{i=1}^{n} \infty_i y_i = 0, \quad 0 \leq \infty_i \leq L \tag{2.2}$$

donde:

- x = los vectores de entrenamiento;
- y = la etiqueta asociada con los vectores de entrenamiento;
- ∞ = los vectores de parámetros del hiperplano clasificador;
- K = una función kernel para medir la distancia entre los vectores de entrenamiento x_i y x_j;
- L = un parámetro de penalización para controlar el número de clasificaciones erróneas.

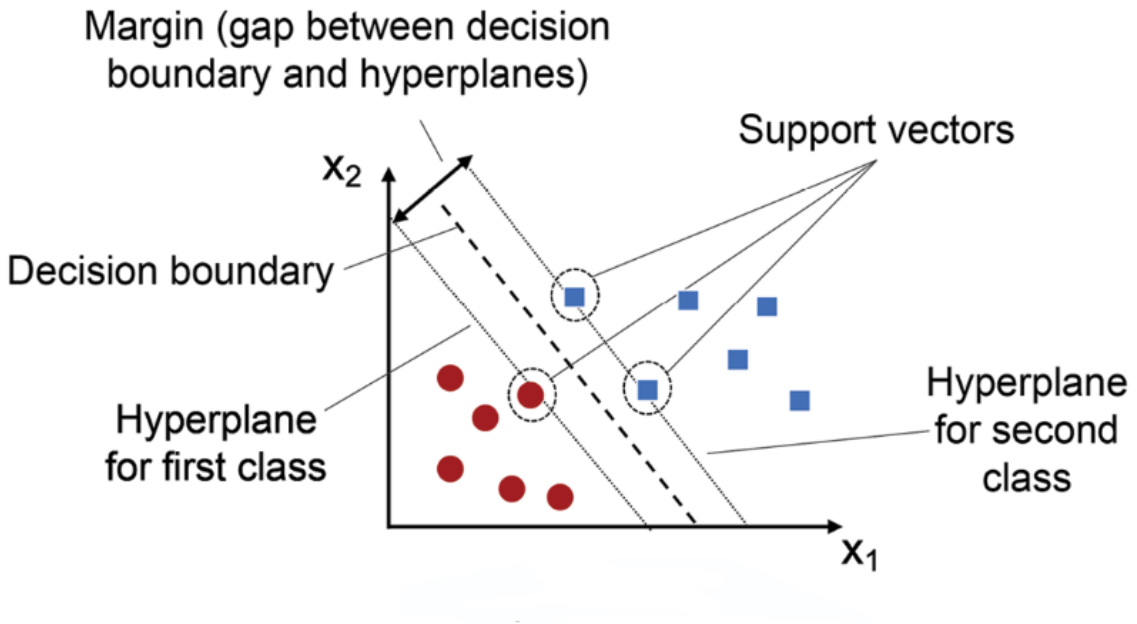


Figure 2.6: Representación esquemática de clasificadores supervisados utilizados en radiómica, las máquinas de vectores de soporte (SVM).

Aunque muchas de las características utilizadas en radiómica, como las texturales y las derivadas de filtros o transformaciones matemáticas, no son recientes y se han empleado desde hace décadas [42], su valor actual radica en la capacidad de analizarlas de forma masiva. La radiómica se fortalece al integrar grandes volúmenes de datos con métodos estadísticos avanzados, aprovechando la experiencia desarrollada en disciplinas -omicas y en el análisis de datos a gran escala, para obtener información útil en el diagnóstico, pronóstico o tratamiento personalizado.

De este modo, la radiómica puede considerarse parte del primer paradigma en la clasificación de imágenes, en el cual tanto la extracción de características como el diseño de algoritmos eran realizados manualmente por expertos. Este enfoque, basado en el conocimiento previo y en descriptores diseñados de forma explícita, fue durante años el estándar en tareas de análisis de imágenes. Sin embargo, este paradigma comenzó a transformarse en 2012, cuando la red neuronal convolucional AlexNet logró un rendimiento sin precedentes en la competencia ImageNet, reduciendo drásticamente el error de clasificación en comparación con métodos tradicionales[43]. Este avance marcó el inicio de la era del aprendizaje profundo (deep learning), en la que los modelos son capaces de aprender automáticamente las representaciones más relevantes directamente desde los datos, sin necesidad de definir manualmente las características.

2.1.6 Características de textura

Las características obtenidas a partir de estadísticas de primer orden derivadas del histograma de una imagen no capturan las relaciones intervoxel. Es en este punto donde entran en juego las características de textura, consideradas de segundo orden, ya que tienen en cuenta las relaciones entre los niveles de gris de cada vóxel y los de sus vecinos. Tal como se mencionó previamente, estas características se agrupan en distintas familias, entre las que se incluyen GLCM, GLSZM, GLDM, entre otras.

Para la mayoría de las familias de características de textura, es necesario realizar un paso

adicional en el preprocesamiento de las imágenes, que consiste en discretizar sus niveles de gris en intervalos conocidos como bins. Este proceso, además de ser un requisito para ciertos métodos de extracción, contribuye a reducir el ruido presente en la imagen.

En lo que sigue de esta sección se explicará como se construyen las matrices GLCM, GLRLM y GLSZM. A partir de estas matrices es posible definir una densidad de probabilidad para calcular las llamadas características de textura, en el apéndice se detallan algunas de las características texturales más relevantes de estas familias, junto a su interpretación en el contexto de la clasificación de lesiones mamarias.

GLCM

Esta familia de características de textura fue propuesta originalmente por Robert M. Haralick, quien demostró en sus trabajos iniciales que eran computacionalmente eficientes y efectivas para la clasificación de imágenes satelitales y aéreas. Desde su introducción, estas características han sido aplicadas exitosamente en una amplia gama de tareas dentro del campo del análisis de imágenes [44].

La matriz de co-ocurrencia de niveles de gris (gray level co-occurrence matrix), GLCM por sus siglas en inglés, es una matriz que representa la distribución de las combinaciones de los niveles de gris de píxeles vecinos, o vóxeles en un volumen 3D, a lo largo de una dirección específica de la imagen.

Matemáticamente, una *GLCM* de tamaño $N_g \times N_g$ se define como $P(i, j \mid \delta, \theta)$, donde el elemento $(i, j)^{\text{th}}$ de la matriz representa el número de veces que la combinación de intensidades i y j ocurre en dos píxeles de la imagen separados por una distancia δ píxeles en la dirección θ. La distancia δ, medida desde el centro del vóxel, se determina utilizando la **norma infinita**. Para $\delta = 1$, esto corresponde a 8 vecinos en 2D, siguiendo los vectores de dirección $(1, 0, 0)$, $(1, 1, 0)$, $(0, 1, 0)$ y $(-1, 1, 0)$, como se puede ver en la figura 2.7 (a) [38].

En la figura 2.7 se ilustra un ejemplo de cómo se calculan los valores de una GLCM, en este

caso vemos cómo la imagen original tiene dos parejas de vecinos (4,1), sombreadas, pues se encuentran a una distancia de un píxel, con $\delta = 1$ y están a lo largo de la dirección horizontal dada por $\theta = 0° = (1, 0)$, por lo que su respectiva entrada en la GLCM $P(i, j \mid 1, 0°)$ ilustrada en 2.7 (c) , es el elemento matricial $(4, 1)$ de valor 2. Análogamente, la pareja de vecinos (1,4) tiene un valor de 2 en el elemento matricial (1,4) de la GLCM en su dirección complementaria, de la figura 2.7 (d).

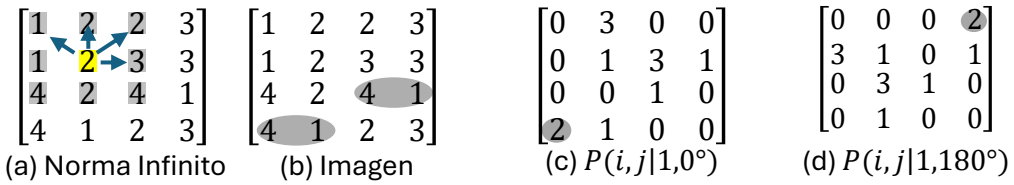


Figure 2.7: (a) Ilustración de la norma infinito en una matriz 2x2, donde vemos los puntos equidistantes con $\delta = 1$ sombreados, con respecto al pixel central resaltado en amarillo, también se ilustran los vectores dirección $(1, 0, 0)$, $(1, 1, 0)$, $(0, 1, 0)$ y $(-1, 1, 0)$. (b) Imagen original con los valores de los niveles de gris de sus pixeles tras un proceso de discretización de bines, GLCM P(i,j) para el vector dirección (1,0) (c),P(i,j) para el vector dirección (-1,0) (d)

GLRLM

Las características texturales GLRM se remontan a la década de los 70. Siguiendo los estudios de Haralick, fueron propuestas por Gallorway nuevas características de esta familia, con el fin de ser aplicadas en la clasificación del mismo conjunto de terrenos utilizado por Haralick. Como resultado, se obtuvieron buenos desempeños que, a su vez, impulsaron y consolidaron el uso de estas características en el campo de clasificación de imágenes [45].

La *Gray Level Co-occurrence Matrix* (*GLCM*) evalúa la co-ocurrencia de niveles de gris en píxeles o vóxeles vecinos, mientras que la *Gray Level Run Length Matrix* (*GLRLM*) cuantifica las carreras de niveles de gris. Una carrera se define como una secuencia de píxeles o vóxeles consecutivos que presentan el mismo valor de nivel de gris, dispuestos en línea recta siguiendo una dirección m o un ángulo θ previamente definidos. La longitud de la carrera corresponde al número de elementos que conforman dicha secuencia. En una matriz *GLRLM* $P(i, j|\theta)$, el

elemento (i, j) representa el número de carreras con nivel de gris i y longitud j que ocurren en la imagen (ROI) siguiendo la dirección especificada [38].

Por ejemplo, en la figura 2.8 (a) se presentan los niveles de gris originales de una imagen. Si construimos la matriz vertical $P(i, j \mid 90°)$ de dimensiones $N_g \times N_r$, donde N_g corresponde al número de niveles de gris y N_r al máximo valor posible de la longitud de una carrera, observamos que la región sombreada en (a) representa una carrera vertical del nivel de gris 2 y de longitud 3. Este caso corresponde al elemento $(2, 3)$ sombreado en la matriz $P(i, j \mid 90°)$ ilustrada en la figura 2.8 (b).

Figure 2.8: Construcción de una GLRLM. Niveles de gris originales (a) y matriz *GLRLM* para la dirección vertical $m = (0, 1)$, donde se ha sombreado el elemento (2,3) que señala 1 carrera de longitud 3 con niveles de gris 2 en la dirección vertical de la imagen original (b).

GLSZM

En el campo del análisis de imágenes, los descriptores de textura basados en matrices de co-ocurrencia *GLCM* y de longitud de carrera *GLRLM* son herramientas consolidadas que han demostrado ser eficaces en diversas tareas de clasificación. Sin embargo, siempre existe la posibilidad de enriquecer los vectores de características para mejorar su poder descriptivo. En un estudio específico sobre la clasificación de núcleos celulares, donde se buscaba diferenciar entre texturas normales (homogéneas) y anormales (heterogéneas), se observó que el uso exclusivo de las características *GLCM* y *GLRLM* no lograba superar el 90 % de exactitud. Para resolver esta limitación, Guillaume et al. introdujeron una nueva familia de descriptores derivados de las GLRLM, denominados **Matriz de Zonas por Tamaño y Nivel de Gris (GLSZM)**. Al combinar estas nuevas características con las ya existentes, el rendimiento del clasificador mejoró significativamente, alcanzando una tasa de éxito del **92.6 %** [46].

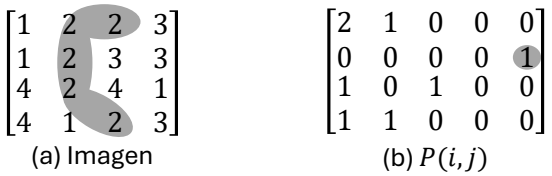


Figure 2.9: Construcción de una GLSZM. Niveles de gris originales (a) y matriz *GLSZM* donde se ha sombreado el elemento (2,5) con valor 1, pues en la imagen original hay un vecindario con tamaño de 5 asociado al nivel de gris 2 (b).

La **Matriz de Zonas por Tamaño y Nivel de Gris (GLSZM)** es un descriptor de textura que caracteriza una imagen mediante la cuantificación de sus zonas de nivel de gris. Este método se basa en el concepto de zona, la cual se define como un conjunto de vóxeles conectados entre sí que poseen una intensidad de gris idéntica. La conectividad entre vóxeles se establece utilizando la norma del infinito, donde la distancia debe ser igual a 1. Esto corresponde a una región de 8 conexiones en un plano 2D.La matriz GLSZM resultante, denotada como $P(i, j)$, se construye de tal manera que cada elemento (i, j) registra el número total de zonas encontradas en la imagen que tienen un nivel de gris i y un tamaño (o área) de j vóxeles [38].

Por ejemplo, en la figura 2.9 (a) se presentan los niveles de gris originales de una imagen. Si construimos la matriz $P(i, j)$ de dimensiones $N_g \times N_z$, donde N_g corresponde al número de niveles de gris y N_z al máximo valor posible de la zona, observamos que la región sombreada en (a) representa una zona del nivel de gris 2 y de tamaño 5. Este caso corresponde al elemento $(2, 5)$ sombreado en la matriz $P(i, j)$ ilustrada en la figura 2.9 (b).

A diferencia de otros métodos como GLCM y GLRLM, la GLSZM posee la ventaja de ser inherentemente independiente de la rotación. Esto significa que solo es necesario calcular una única matriz para toda la región de interés (ROI), en lugar de promediar matrices de distintas direcciones.

Si bien las características radiómicas proporcionan descriptores cuantitativos explícitos de las imágenes, su diseño manual requiere conocimiento experto del dominio. En contraste,

el aprendizaje profundo, que exploraremos en la siguiente sección, permite que los modelos aprendan automáticamente representaciones jerárquicas de características directamente desde los datos, representando un cambio paradigmático en el análisis de imágenes médicas.

2.2 Redes neuronales

2.2.1 Antecedentes históricos

Una de las ventajas principales del ser humano ante otras especies es su capacidad de aprender y transmitir ese conocimiento de generación en generación. Si bien la comprensión de cómo aprendemos ha sido un largo debate en la filosofía, a nivel biológico sabemos que los procesos cognitivos relacionados con el aprendizaje se dan en las neuronas de nuestro cerebro. Fue precisamente el intento de modelar el funcionamiento de estas células biológicas lo que dio inicio al campo de las redes neuronales artificiales.

El primer paso fundamental en esta dirección lo dieron McCulloch y Pitts en 1943, quienes propusieron un modelo matemático de las redes neuronales biológicas que sentó las bases teóricas de la neurona artificial. Sobre este fundamento, a finales de la década de 1950, Frank Rosenblatt desarrolló el Perceptrón, un modelo que representó un avance crucial al incorporar una regla de aprendizaje para resolver problemas de reconocimiento de patrones [47, 48]. Sin embargo, en 1969, Marvin Minsky y Seymour Papert publicaron *Perceptrons*, una obra crítica que demostró las limitaciones del perceptrón simple para resolver problemas no linealmente separables, como el XOR. Este señalamiento generó un escepticismo generalizado y condujo a una disminución drástica del interés y la financiación en el área, fenómeno conocido como el *invierno de las redes neuronales* [49].

El campo experimentó un resurgimiento en la década de 1980 gracias a varios factores: la introducción de arquitecturas multicapa, el aumento de la capacidad computacional disponible, y sobre todo, la popularización del algoritmo de retropropagación (*backpropagation*), descrito

por Rumelhart, Hinton y Williams en 1986, que permitió entrenar de manera eficiente redes con múltiples capas ocultas y abordar problemas de mayor complejidad [50]. Asimismo, contribuciones como el modelo Adaline de Widrow y Hoff (1960) mantuvieron viva la investigación en aprendizaje adaptativo durante los años intermedios [51].

2.2.2 Perceptron multicapa

El diseño del Perceptrón Multicapa (MLP) está directamente inspirado en el funcionamiento de las redes neuronales biológicas, un concepto que se remonta a los primeros modelos computacionales de la neurona [47]. En una neurona biológica, las señales electroquímicas son recibidas por las dendritas, se integran en el soma (cuerpo celular) y, si la estimulación acumulada supera un cierto umbral, se dispara un impulso que se transmite a través del axón hacia otras neuronas.

El perceptrón artificial emula este flujo. Recibe un conjunto de entradas, las procesa mediante una suma ponderada y aplica una función de activación, que simula el disparo del impulso nervioso para determinar la señal de salida [48]. Esta estructura básica es el pilar de las redes neuronales profundas modernas [52]. En la Figura 2.10 se ilustra esta comparación.

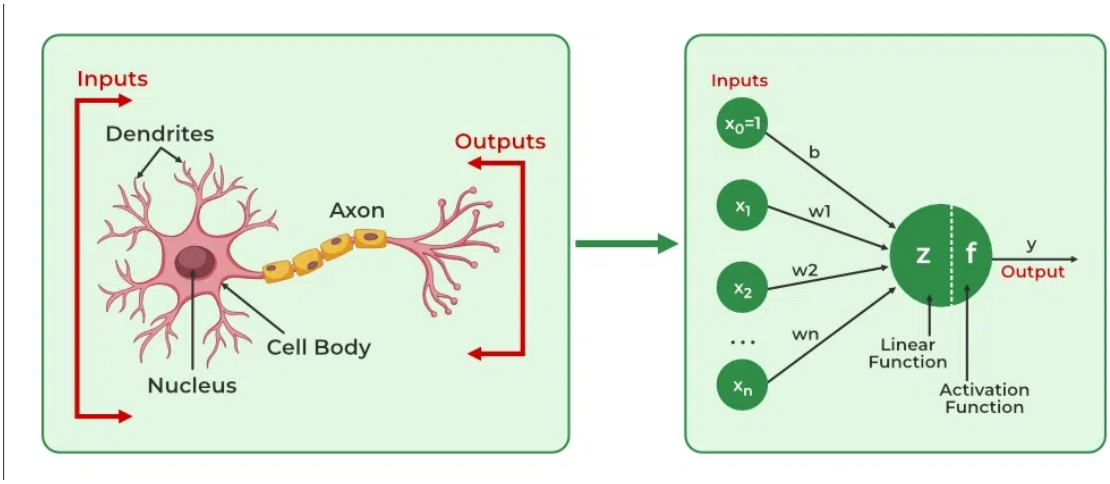


Figure 2.10: Comparación esquemática entre una neurona biológica (izquierda) y su análogo artificial, el perceptrón (derecha).

Matemáticamente, la integración de señales que realiza la neurona artificial j en una MLP

se representa como:

$$\sum_{i=1}^{q} w_{i,j} x_i + b_j \tag{2.3}$$

La interpretación biológica de cada componente es clave para entender el modelo [53]:

- x_i (Entradas): Son análogas a la frecuencia de los impulsos nerviosos que una neurona recibe en sus dendritas.
- $w_{i,j}$ (Pesos Sinápticos): Simulan la fuerza o eficiencia de la conexión sináptica. Un peso positivo ($w_{i,j} > 0$) equivale a una sinapsis excitatoria, mientras que un peso negativo ($w_{i,j} < 0$) representa una sinapsis inhibitoria.
- b_j (Sesgo o Bias): Es análogo al umbral de activación intrínseco de la neurona, determinando la cantidad de estímulo neto necesario para que se active.

El resultado de la Ecuación 2.3 es el valor que se introduce en la función de activación para obtener la salida final de la neurona, la cual se transmitirá a la siguiente neurona. En la próxima sección hablaremos sobre cómo se suelen organizar las neuronas.

2.2.3 Arquitectura de una red neuronal

Las neuronas artificiales se organizan en una estructura de capas secuenciales, donde cada capa consiste en un grupo de neuronas que procesan la información antes de pasarla a la siguiente. La arquitectura de una red neuronal se compone fundamentalmente de tres tipos de capas:

- Capa de Entrada (Input Layer): Actúa como el portal de la red. Su función no es procesar información, sino recibir los datos iniciales —como los píxeles de una imagen o las palabras de un texto— y transmitirlos sin modificación hacia la primera capa oculta.
- Capas Ocultas (Hidden Layers): Son las capas intermedias que se encuentran entre la entrada y la salida. Aquí ocurre el procesamiento principal. Las neuronas de estas

capas reciben la información de la capa anterior, aplican transformaciones y extraen características cada vez más complejas y abstractas de los datos.

- Capa de Salida (Output Layer): Es la capa final de la red. Consolida el procesamiento realizado en las capas ocultas para producir el resultado final de la tarea, que puede ser una categoría de clasificación, un valor numérico de una predicción, entre otros.

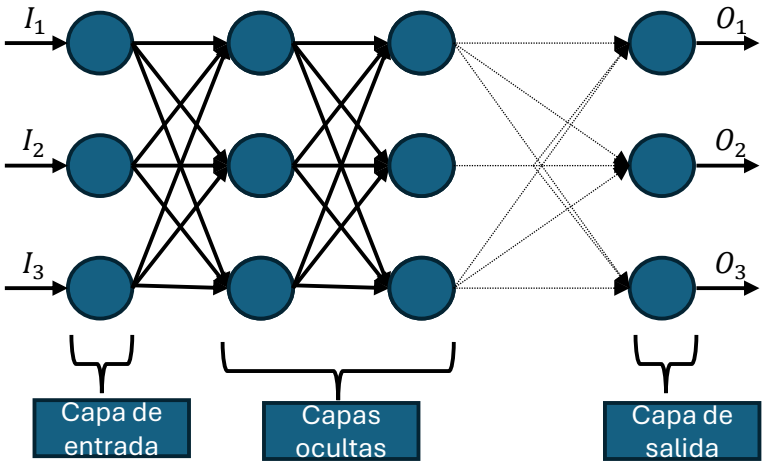


Figure 2.11: Esquema de la organizacion de una red . En la capa de entrada entran las componentes del vector input, luego en las capas ocultas son hacen las sumas ponderadas de los pesos y se transmiten la información hacia adelante, finalmente la capa de salida da una predicción en el contexto de la clasificación.

La arquitectura específica de una red —es decir, el número de capas ocultas y la cantidad de neuronas en cada una— no es fija y se diseña en función de la complejidad del problema a resolver. Mientras que tareas sencillas pueden manejarse con una única capa oculta, los problemas más desafiantes requieren arquitecturas profundas con múltiples capas y millones de neuronas.

Aumentar la profundidad de una red (añadir más capas) potencia su capacidad para aprender patrones más sutiles y jerárquicos en los datos, lo que a menudo mejora el rendimiento en tareas complejas. No obstante, una profundidad excesiva puede ser contraproducente,

ya que incrementa el riesgo de sobreajuste (*overfitting*). Este fenómeno ocurre cuando el modelo se memoriza los datos de entrenamiento tan específicamente que pierde su capacidad de generalizar y funcionar correctamente con datos nuevos y nunca antes vistos.

2.2.4 Funciones de activación

El rol de una función de activación es introducir no linealidad en la salida de una neurona. De lo contrario, la operación de la neurona se limitaría únicamente a la transformación lineal que es definida por sus pesos y su sesgo (*bias*).

Por esta razón, las funciones de activación juegan un papel indispensable en el aprendizaje automático. Estas proveen la capacidad necesaria para modelar fenómenos del mundo real, los cuales, a menudo, se caracterizan por tener relaciones complejas y no lineales entre sus variables.

Entre las funciones de activación más comunes tenemos:

- Función Escalonada: La función escalonada o de Heaviside, denotada por $H(x)$ o $\theta(x)$, es una función discontinua que produce una salida binaria: 1 para argumentos positivos o nulos, y 0 para negativos.

$$H(x) := \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

  Su principal desventaja es que su derivada es cero en casi todos los puntos y no está definida en $x = 0$. Esto la hace incompatible con algoritmos de entrenamiento basados en el descenso de gradiente, como la retropropagación.

- Función Sigmoide: La función sigmoide, también conocida como función logística, mapea cualquier valor de entrada real a un rango de salida acotado entre (0, 1). Se define de la siguiente forma:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Su característica forma de 'S' satura las entradas de gran magnitud, es decir, mapea valores muy positivos a 1 y valores muy negativos a 0. Si bien su salida es interpretable como una probabilidad, esta saturación es problemática para el entrenamiento de redes profundas. Cuando la entrada a la función es muy grande o muy pequeña, la curva de la sigmoide se vuelve casi plana, lo que significa que su derivada (gradiente) se aproxima a cero.

Durante el entrenamiento mediante retropropagación (*backpropagation*), los gradientes de las capas más profundas se calculan multiplicando los gradientes de las capas superiores. Si las neuronas con activación sigmoide saturan, sus pequeños gradientes se multiplican en cadena, provocando que el gradiente final que llega a las primeras capas del modelo se "desvanezca" o se vuelva extremadamente pequeño. Este fenómeno, conocido como el problema del desvanecimiento del gradiente (*vanishing gradient problem*), impide que los pesos de las primeras capas se actualicen de manera efectiva, deteniendo o ralentizando drásticamente el aprendizaje de la red [54, 55].

- Función Tangente Hiperbólica (Tanh): Esta función presenta una forma sigmoidal similar a la logística, pero su rango de salida está centrado en cero, abarcando el intervalo (-1, 1).

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

  Esta simetría respecto al origen a menudo conduce a una convergencia más rápida durante el entrenamiento. Además, su gradiente es más pronunciado que el de la sigmoide, lo que la convierte en una opción preferida para mitigar el problema del desvanecimiento del gradiente (*vanishing gradient*).

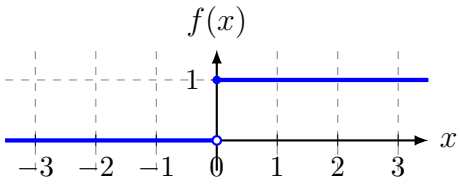
- Función ReLU: La Unidad Lineal Rectificada (ReLU) es una función de activación que

calcula la parte positiva de su argumento, comportándose como una función rampa.

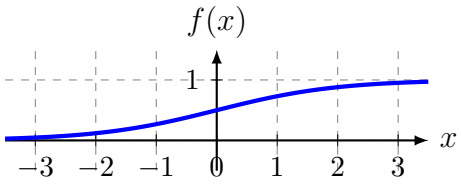
$$\mathrm{ReLU}(x) = x^+ = \max(0, x) = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

Según Glorot et al. (2011), sus principales ventajas frente a las funciones sigmoide y tanh son: una mayor similitud con la respuesta de las neuronas biológicas (esparcidad de activación), una mitigación efectiva del problema del desvanecimiento del gradiente para valores positivos, y una eficiencia computacional superior. Estas ventajas convierten a la funcion ReLU en una de las funciones de activación mas utilizadas en la actualidad. [55].

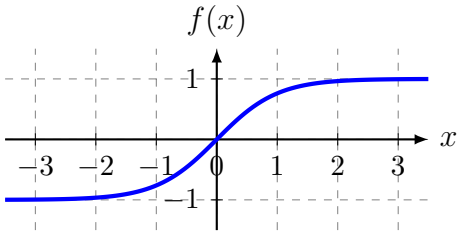
Sin embargo, ReLU no está exenta de limitaciones. Su principal desventaja es el problema de la "neurona muerta" (*dying ReLU problem*). Si una neurona ReLU recibe una entrada consistentemente negativa, su salida será siempre cero. En consecuencia, su gradiente también será cero, lo que impide que los pesos asociados a ella se actualicen durante la retropropagación. La neurona queda "atascada" en un estado inactivo y deja de contribuir al aprendizaje del modelo [56]. Adicionalmente, al no estar acotada superiormente, las activaciones pueden crecer de forma descontrolada, lo que en redes muy profundas puede llevar a problemas de inestabilidad numérica o al fenómeno conocido como "explosión del gradiente" (*exploding gradients*).



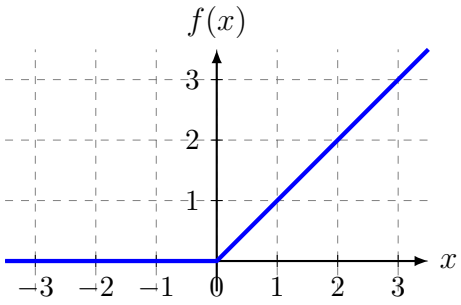
(a) Función Escalonada (Heaviside)



(b) Función Sigmoide



(c) Función Tangente Hiperbólica (Tanh)



(d) Función ReLU

Figure 2.12: Gráficas de funciones de activación comunes.

Función	Ventajas Clave	Desventajas Clave
Escalonada	Simplicidad conceptual	Gradiente nulo (incompatible con backprop)
Sigmoide	Salida probabilística (0, 1)	Vanishing gradient, no centrada en cero
Tanh	Centrada en cero, mejor gradiente	Aún sufre de vanishing gradient
ReLU	Rápida, no satura (para x > 0)	Dying ReLU, no centrada en cero, no acotada

Table 2.3: Resumen de ventajas y desventajas de funciones de activación.

2.2.5 Backpropagation

El algoritmo de retropropagación (*Backpropagation*) es la técnica fundamental para el entrenamiento de redes neuronales, diseñada para ajustar los pesos de la red de manera iterativa a partir del error de predicción.

El proceso de entrenamiento con *Backpropagation* se divide en dos fases principales que se repiten en cada iteración: la propagación hacia adelante y la propagación hacia atrás. En la propagación hacia adelante (*forward pass*), los datos de entrada atraviesan la red desde la primera capa hasta la última para generar una predicción o clasificación.

Para ilustrar cómo se calculan la predicción y el error en una tarea de clasificación, a continuación se describen la función Softmax, comúnmente utilizada en la capa de salida, y su función de pérdida asociada.

La función Softmax toma un vector de puntuaciones brutas (conocidas como *logits*), z, y lo transforma en una distribución de probabilidad. Cada componente j del vector de salida, $f_j(z)$, representa la probabilidad de que la entrada pertenezca a la clase j. Una propiedad fundamental de esta función es que la suma de las probabilidades de todas las categorías siempre es igual a 1. Su forma funcional se define en la Ecuación 2.4:

$$f_j(z) = \frac{e^{z_j}}{\sum_k e^{z_k}} \tag{2.4}$$

Para cuantificar el error de la predicción generada por Softmax, se utiliza la función de pérdida de Entropía Cruzada (*Cross-entropy loss*). Esta función mide qué tan bien la distribución de probabilidad predicha por el modelo se alinea con la clase verdadera. La fórmula, que combina eficientemente la función Softmax y el logaritmo negativo, se expresa de la siguiente manera:

$$L_i = -\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right) \tag{2.5}$$

Donde L_i es la pérdida para una única muestra de entrenamiento, y f_{y_i} es el logit (la salida bruta de la red) correspondiente a la clase correcta y_i. El valor de esta pérdida es el que se propaga hacia atrás para ajustar los pesos de la red [57].

Posteriormente, en la propagación hacia atrás (*backward pass*), el error calculado se propaga en sentido inverso, desde la capa de salida hacia la de entrada. Para determinar la contribución de cada peso al error total, el algoritmo se fundamenta en la regla de la cadena del cálculo diferencial.

Esta regla permite calcular el gradiente de la función de pérdida (E) con respecto a cada peso (w_{ij}) y sesgo (b_j) de la red, incluso aquellos en las capas más profundas. De forma intuitiva, la regla de la cadena descompone el impacto de un peso en el error final en una serie de influencias locales: el impacto del peso en la salida de su neurona, el impacto de

esa neurona en la siguiente capa, y así sucesivamente, hasta llegar a la capa de salida. Este cálculo se realiza de manera eficiente propagando el gradiente hacia atrás, desde la última capa hasta la primera [52].

Finalmente, los pesos y sesgos se actualizan en la dirección opuesta al gradiente para minimizar el error. La magnitud de esta actualización es modulada por un hiperparámetro conocido como tasa de aprendizaje (η):

$$w_{ij} \leftarrow w_{ij} - \eta \frac{\partial E}{\partial w_{ij}} \tag{2.6}$$

$$b_j \leftarrow b_j - \eta \frac{\partial E}{\partial b_j} \tag{2.7}$$

Este proceso iterativo de propagación hacia adelante y hacia atrás es el que permite a la red "aprender" a partir de los datos.

2.3 Redes neuronales convolucionales (CNN)

Las *CNN*, son un tipo de modelo de aprendizaje profundo especialmente diseñado para procesar datos con estructura en forma de cuadrícula, como las imágenes digitales. Estas redes se inspiran en la organización jerárquica del córtex visual de los animales, el cual está compuesto por neuronas especializadas en detectar distintos tipos de estímulos visuales. De manera análoga, las *CNN* están diseñadas para aprender automáticamente jerarquías espaciales de características, avanzando desde patrones simples como bordes, esquinas o texturas, hasta representaciones complejas que pueden capturar objetos completos o contextos más abstractos [58, 59].

La arquitectura básica de una *CNN* está formada por tres tipos principales de capas: capas de convolución, capas de agrupamiento (pooling) y capas completamente conectadas (fully connected), como se ilustra en la figura 2.13.

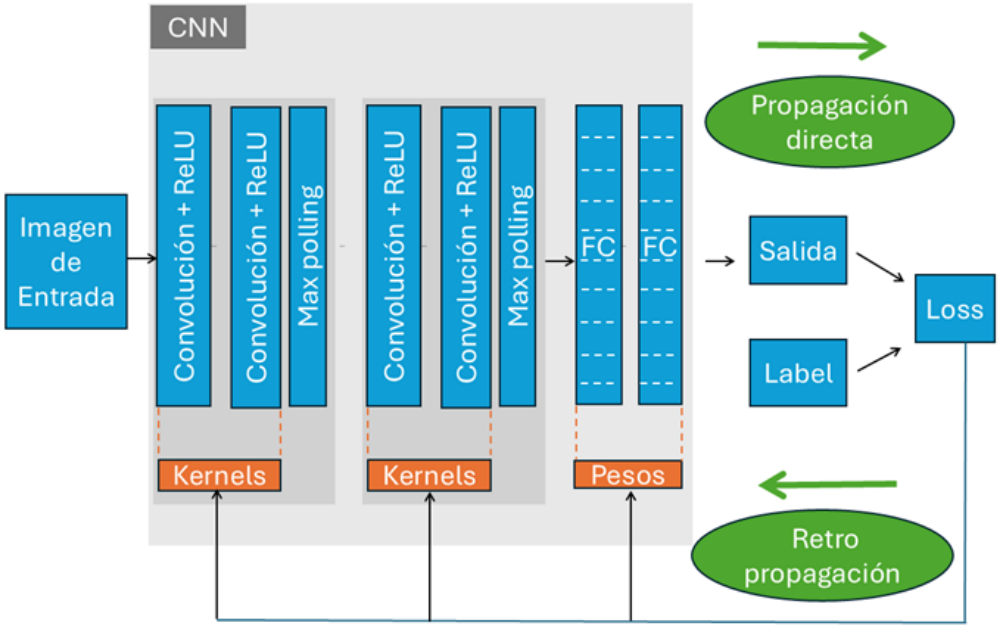


Figure 2.13: Esquema general del funcionamiento de una *CNN*, mostrando las etapas de propagación directa y retropropagación durante el proceso de entrenamiento. La imagen de entrada se procesa a través de capas de convolución, activación ReLU, Max Pooling, seguida de capas completamente conectadas (FC), ajustando los pesos mediante el cálculo del error (Loss) para optimizar el modelo.

2.3.1 Capas de convolución

En el procesamiento de imágenes estas se representan como matrices bidimensionales de valores numéricos correspondientes a la intensidad de cada píxel, las *CNN* aplican filtros o *kernels* entrenables que recorren la imagen con una ventana deslizante. Estos filtros permiten detectar patrones específicos sin importar su ubicación exacta, otorgando a las redes una propiedad conocida como invariancia traslacional. Cada elemento de estos kernels corresponde a un peso, el cual es iniciado aleatoriamente, pero mediante el *backpropagation* estos pesos se actualizan para detectar mejor las características.

La Figura 2.14 muestra un kernel de 3x3 diseñado para la ilustrar de forma pedagógica la detección de bordes con orientación diagonal. Cuando este kernel (K) se aplica a una imagen de entrada (I), actúa como un detector de características.

En el contexto de las redes neuronales, esta operación es comúnmente llamada convolución.

Sin embargo, es técnicamente una correlación cruzada (*cross-correlation*), ya que el kernel no se invierte como ocurriría en una convolución matemática estricta [52]. Esta operación desliza el kernel sobre la imagen y, en cada posición, calcula la suma ponderada de los píxeles subyacentes.

Para manejar la aplicación del kernel en los bordes de la imagen y controlar las dimensiones del mapa de características resultante, se utilizan dos hiperparámetros:

- Relleno (Padding): Consiste en añadir un borde de píxeles (generalmente con valor cero) alrededor de la imagen de entrada. Esto permite que el kernel pueda operar sobre los píxeles que se encuentran en los bordes originales de la imagen y, además, ayuda a preservar las dimensiones espaciales de entrada.

- Paso (Stride): Define el número de píxeles que el kernel se desplaza en cada paso sobre la imagen. Un *stride* de 1 mueve el kernel un píxel a la vez, mientras que un valor mayor produce un mapa de características de menor tamaño.

La operación se define por la Ecuación 2.8. Para entenderla, es crucial recordar que los índices (i,j) se refieren a la posición del píxel en el **mapa de características de salida (S)**. El *stride* (S_h, S_w) mapea estas coordenadas de salida a la esquina superior izquierda de la región en la imagen de entrada (I) donde se aplica el kernel.

- Para calcular el primer píxel de salida $S(0,0)$, el kernel se posiciona en la esquina ‘(0,0)‘ de la imagen de entrada.

- Para calcular el siguiente píxel a la derecha, $S(0,1)$, el kernel se desplaza S_w píxeles a la derecha en la entrada, posicionándose en ‘(0, S_w)‘.

- Generalizando, para calcular el píxel de salida $S(i,j)$, el kernel se posiciona en la esquina ‘$(i \cdot S_h,\ j \cdot S_w)$‘ de la imagen de entrada.

Una vez posicionado el kernel, la suma se realiza sobre los índices del kernel (m,n) para cubrir la región correspondiente de la imagen. Esto nos lleva a la siguiente fórmula:

$$S(i,j) = \sum_m \sum_n I(i \cdot S_h + m, j \cdot S_w + n) K(m,n) \tag{2.8}$$

donde S_h y S_w son los pasos (*strides*) vertical y horizontal.

El resultado de esta convolución es un mapa de características que resalta las áreas de la imagen que contienen el borde diagonal.En la figura 2.15 se puede observar un ejemplo del mapa de características en un estudio de COVID-19. Posteriormente, a este mapa se le aplica la función de activación ReLU.

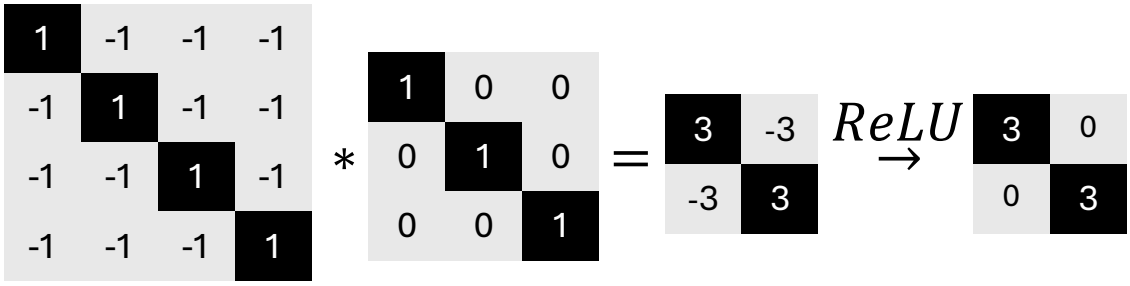


Figure 2.14: Ejemplo de un Kernel 3x3 especializado en la detección de bordes diagonales y como actúan la función de convolución y activación.

La función ReLU cumple el rol de introducir no linealidad y filtrar el mapa, suprimiendo todas las activaciones negativas (que no se corresponden con la característica buscada) y manteniendo únicamente los valores positivos. De este modo, en el mapa de activación final, la magnitud de los valores es directamente proporcional a la presencia de la característica: los valores altos indican una fuerte correspondencia con el borde diagonal, mientras que los ceros indican la ausencia de dicha característica en esa región de la imagen.

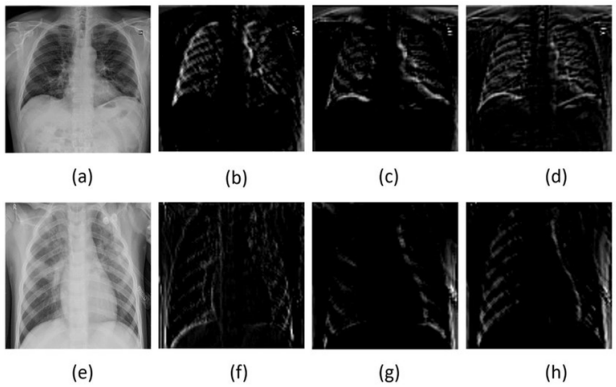


Figure 2.15: Visualización de mapas de características: (a) Caso positivo de COVID-19 y sus correspondientes mapas de características (b)-(D); y (e) Caso positivo de neumonía y sus correspondientes mapas de características (f)-(g)[60]

2.3.2 Capas de pooling

Una función de *pooling* reemplaza la salida de una capa de convolución en una ubicación específica por un resumen estadístico de las salidas cercanas. Una de las variantes más utilizadas es el Max Pooling, la cual extrae el valor máximo dentro de una vecindad rectangular, como se puede observar en la Figura 2.16.

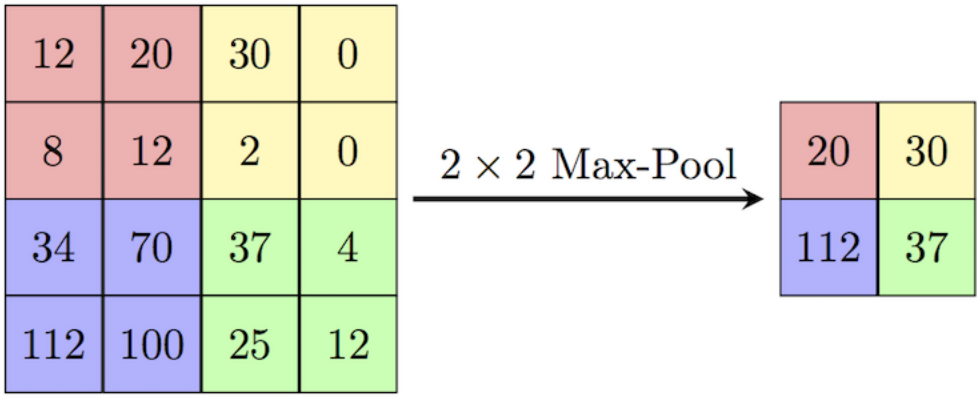


Figure 2.16: Ejemplo de una operación de Max Pooling.

Existen otros tipos de *pooling*, como el *Average Pooling*, que en su lugar calcula el promedio de la región rectangular. En cualquier caso, el *pooling* ayuda a que la representación de características sea aproximadamente invariante ante pequeñas traslaciones de la entrada. Esta propiedad resulta útil cuando el interés principal es detectar la presencia de una característica,

más que su ubicación exacta. Por ejemplo, para determinar si una imagen contiene un rostro, basta con saber que hay un ojo en el lado izquierdo y otro en el lado derecho, sin necesidad de conocer su posición con precisión a nivel de píxel [52].

Sin embargo, en otros contextos, preservar la localización de una característica es fundamental. Un caso de esto es la detección de una esquina definida por la intersección de dos bordes; para esta tarea, es necesario conservar la ubicación de dichos bordes con suficiente detalle para verificar que efectivamente se encuentran.

Además, la reducción de tamaño que realiza la función de *pooling* ayuda a disminuir los costos computacionales y de memoria durante el entrenamiento de la red.

2.3.3 Capas completamente conectadas

Las capas completamente conectadas (*Fully Connected Layers*), ubicadas generalmente al final de una arquitectura de CNN, corresponden al modelo de perceptrón multicapa descrito en la sección anterior. Su función principal es tomar las características de alto nivel extraídas por las capas convolucionales y de *pooling* para realizar la tarea de clasificación final.

La entrada a esta capa es el mapa de características resultante de la última capa convolucional, el cual se somete a una operación de aplanamiento (*flattening*). Este proceso convierte la matriz multidimensional de características en un único vector.

Una vez que las características se encuentran en formato vectorial, se realiza la clasificación. Durante la propagación hacia adelante (*forward propagation*), este vector atraviesa una o más capas completamente conectadas para generar el vector de salida con las puntuaciones de cada clase.

Durante la retropropagación (*backpropagation*), el error de clasificación se utiliza para ajustar los pesos de la red. Es crucial destacar que este ajuste no solo ocurre en las neuronas de la sección completamente conectada, sino que el gradiente del error se propaga hacia atrás a través de toda la red para actualizar también los pesos de los kernels en las capas convolucionales.

2.4 Aprendizaje por transferencia

El aprendizaje por transferencia (*transfer learning*) es una estrategia de aprendizaje automático que se ha consolidado como una de las técnicas más efectivas para el análisis de imágenes médicas, especialmente en el contexto de las redes neuronales profundas (DNNs). Su premisa fundamental consiste en aprovechar el conocimiento adquirido por un modelo al ser entrenado en una tarea o dominio de datos, para luego aplicarlo a una tarea diferente pero relacionada [61].

Fundamento Teórico: La Generalidad de las Características Aprendidas

El éxito del *Transfer Learning* en el dominio de la visión por computador se fundamenta en la naturaleza jerárquica del aprendizaje en las Redes Neuronales Convolucionales (CNNs). Cuando una red se entrena en un conjunto de datos masivo y diverso, como ImageNet (que contiene millones de imágenes de objetos cotidianos), aprende a reconocer un vasto repertorio de características visuales de forma jerárquica [62].

- Capas Iniciales: Las primeras capas de la red aprenden a detectar características muy simples y genéricas, como bordes, esquinas, texturas y gradientes de color. Estas características son universales para casi cualquier tarea de visión.
- Capas Intermedias: A medida que la información avanza, las capas intermedias combinan estas características simples para reconocer patrones más complejos, como formas, partes de objetos (ojos, ruedas) o texturas elaboradas.
- Capas Finales: Las últimas capas aprenden a identificar conceptos de alto nivel, específicos de la tarea original (por ejemplo, “perro”, “coche”, “gato”).

La hipótesis central del *Transfer Learning* es que las características de bajo y medio nivel aprendidas en un dominio general (como ImageNet) son lo suficientemente robustas y genéricas como para ser útiles en un dominio especializado, como las imágenes médicas. Aunque una

radiografía y la foto de un gato son visualmente distintas, ambas están compuestas por los mismos elementos fundamentales: bordes, texturas y formas.

Aplicación en el Diagnóstico por Imagen Médica

El principal desafío en el campo de la imagenología médica es la escasez de datos etiquetados. Crear grandes bases de datos requiere de un enorme esfuerzo, tiempo y costo, además de la experticia de radiólogos para realizar las anotaciones, sin mencionar las estrictas regulaciones de privacidad. El *Transfer Learning* aborda directamente este problema de la siguiente manera: en lugar de entrenar una red neuronal desde cero (con pesos inicializados aleatoriamente), lo que requeriría cientos de miles de imágenes, se utiliza una red pre-entrenada como punto de partida.

Existen dos estrategias principales para implementar esta técnica:

(1) Extracción de Características (Feature Extraction): En este enfoque, se utiliza la base convolucional de la red pre-entrenada (por ejemplo, las capas convolucionales de VGG16 o ResNet50) como un extractor de características fijo. Se eliminan las capas finales de clasificación originales y se añade una nueva cabeza clasificadora, usualmente compuesta por unas pocas capas densas. Durante el entrenamiento, solo se actualizan los pesos de esta nueva cabeza clasificadora, mientras que el resto de la red permanece "congelado". Este método es ideal cuando se dispone de un conjunto de datos médicos muy pequeño.

(2) Ajuste Fino (Fine-Tuning): Esta estrategia es una extensión de la anterior. Se comienza cargando los pesos pre-entrenados en toda la red, pero en lugar de congelar la base convolucional, se permite que una parte de ella (generalmente las capas superiores) también se entrene con los nuevos datos médicos. El entrenamiento se realiza con una tasa de aprendizaje muy baja para no modificar drásticamente los pesos pre-entrenados, sino simplemente "ajustarlos" a las sutilezas del nuevo dominio. Este enfoque suele dar mejores resultados cuando se tiene una cantidad moderada de datos y permite que el modelo adapte sus características de alto nivel al contexto médico.

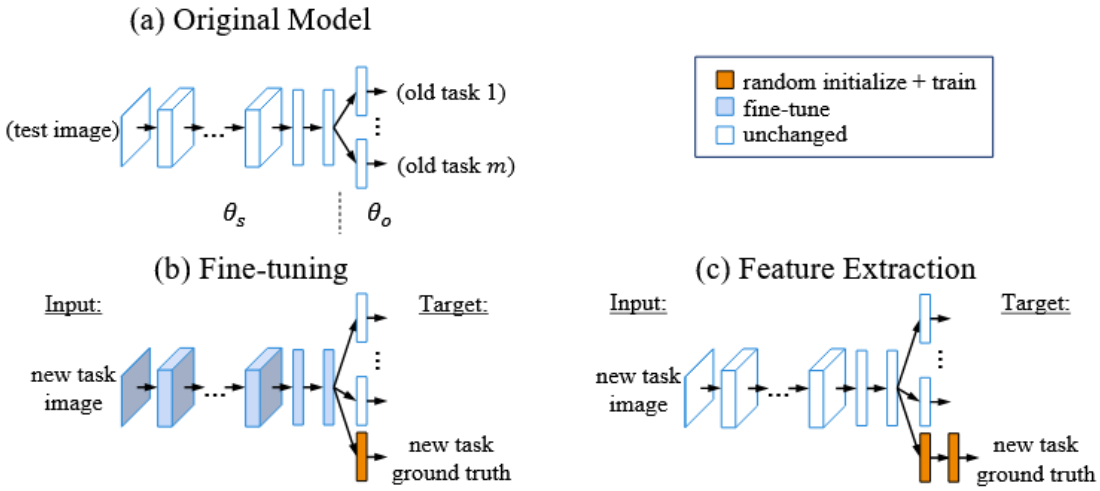


Figure 2.17: Ilustración de la diferencia entre *fine-tuning* y *feature extraction* [63]

Al transferir el conocimiento adquirido de un dominio con abundancia de datos, los modelos pueden alcanzar un alto rendimiento con un número limitado de imágenes clínicas, reduciendo significativamente el tiempo de entrenamiento y la necesidad de recursos computacionales, y mejorando la capacidad de generalización del clasificador final [28, 16].

2.5 Tecnicas de ensamble

2.5.1 Métodos de Votación

El método de votación (*voting*) es una técnica de ensamble que combina las predicciones de múltiples modelos base, entrenados sobre el mismo conjunto de datos. El principio fundamental es agregar los resultados individuales para obtener una predicción final más robusta y precisa, siguiendo un criterio de mayoría. Aunque es aplicable tanto a problemas de regresión como de clasificación, en este contexto se analiza su uso para este último [64].

Dependiendo del tipo de salida que generan los modelos base, la votación para clasificación se divide en dos categorías principales: votación dura y votación ponderada. La figura 2.18 ejemplifica la diferencia entre estas dos.

- Votación Dura (*Hard Voting*): En este esquema, cada modelo del ensamble emite un

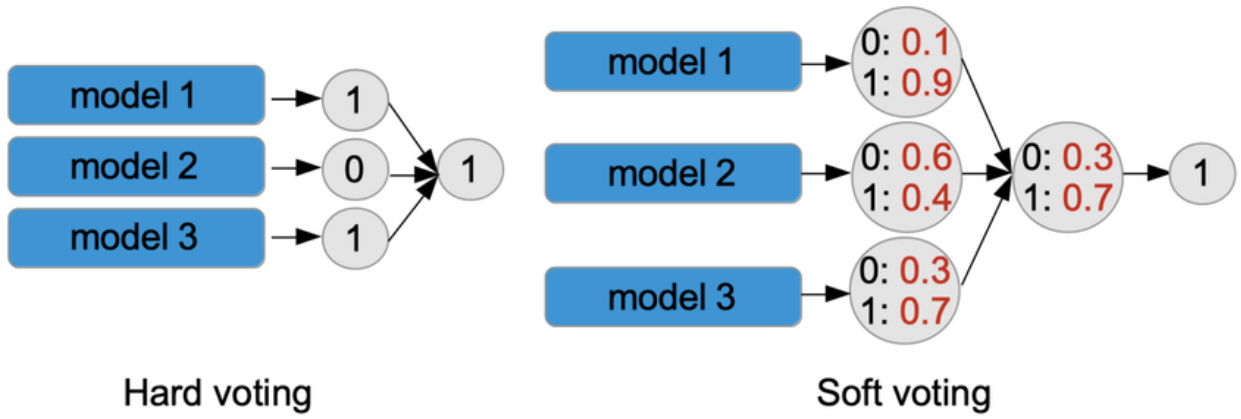


Figure 2.18: Esta imagen muestra a la izquierda un ensamble de hard voting donde se escoje la clase con mayor frecuencia, mientras que la derecha muestra un soft voting donde se tiene en cuenta el promedio de las probabilidades de cada clase

voto por una única etiqueta de clase. La predicción final del conjunto es simplemente la clase que recibe el mayor número de votos (es decir, la moda de las predicciones). Este método es aplicable a cualquier clasificador que pueda generar una predicción de clase discreta.

- Votación Ponderada (*Soft Voting*): Este enfoque, a menudo más potente, utiliza las probabilidades de pertenencia a cada clase que predicen los modelos base. En lugar de un voto único, cada modelo aporta un vector de probabilidades. El ensamble calcula el promedio de estas probabilidades para cada clase, y la predicción final es la clase con la probabilidad promedio más alta. La votación ponderada requiere modelos capaces de generar estas predicciones de probabilidad, o en su defecto, una puntuación análoga a la probabilidad, como es el caso de las Máquinas de Vectores de Soporte (SVM).

2.6 Stacking

El *stacking*, o apilamiento, es una técnica de aprendizaje por conjuntos (ensemble learning) cuyo objetivo es mejorar la capacidad predictiva combinando las predicciones de múltiples modelos. El resultado es un "modelo apilado" final que busca ser más robusto y preciso que

cualquiera de sus componentes individuales [65].

Arquitectura y Proceso

La arquitectura del stacking se organiza en dos niveles. Primero, los *modelos base* (Nivel 0), como pueden ser árboles de decisión o regresión logística, se entrenan sobre los datos originales. A continuación, un *meta-modelo* (Nivel 1), que suele ser un modelo simple, aprende a combinar de manera óptima las predicciones generadas por los modelos base para producir el resultado final.

El proceso consiste en entrenar los modelos base, generar predicciones con ellos sobre un conjunto de validación, y usar estas predicciones como nuevas características para entrenar al meta-modelo. Para inferir sobre datos nuevos, las predicciones de los modelos base se introducen en el meta-modelo para obtener la predicción definitiva.

Ventajas y Limitaciones

Las principales ventajas del stacking son la mejora en el rendimiento predictivo, la capacidad de combinar algoritmos diversos aprovechando sus fortalezas, la reducción del sobreajuste (con una validación cruzada adecuada) y una alta flexibilidad en su configuración.

No obstante, esta técnica también presenta limitaciones notables. Su implementación es más compleja que la de otros ensambles como Bagging o Boosting, su tiempo de entrenamiento es elevado debido al coste computacional, y la estructura de múltiples capas dificulta la interpretación del modelo final (siendo considerado un modelo de "caja negra"). Además, requiere una cantidad considerable de datos para funcionar de manera efectiva.

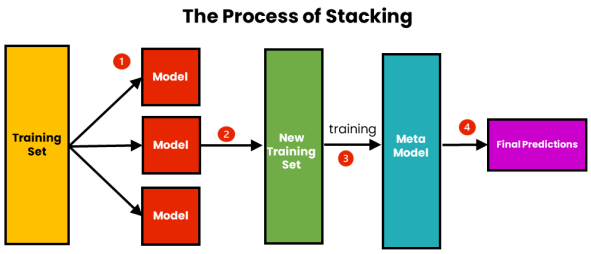


Figure 2.19: Ilustracion de un modelo de stacking

2.7 Métricas

Los modelos previamente descritos son evaluados en su desempeño utilizando un conjunto de métricas ampliamente aceptadas en la literatura biomédica. Para definir estas métricas, primero se consideran los cuatro posibles resultados de una clasificación binaria, usualmente representados en una matriz de confusión:

- Verdaderos Positivos (VP o TP): Casos positivos que fueron clasificados correctamente como positivos.
- Verdaderos Negativos (VN o TN): Casos negativos que fueron clasificados correctamente como negativos.
- Falsos Positivos (FP): Casos negativos que fueron clasificados incorrectamente como positivos (Error Tipo I).
- Falsos Negativos (FN): Casos positivos que fueron clasificados incorrectamente como negativos (Error Tipo II).

A partir de estos resultados se construye la matriz de confusión:

A partir de estos valores, se definen las siguientes métricas:

Sensibilidad La sensibilidad (*recall* o Tasa de Verdaderos Positivos) refleja la capacidad del modelo para identificar correctamente los casos positivos, es decir, aquellos pacientes con

Figure 2.20: Matriz de confusión donde se puede apreciar la distribución de los VP,VN,FP,FN

la condición de interés.

$$\text{Sensibilidad} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

Especificidad La especificidad indica la habilidad del modelo para reconocer correctamente los casos negativos, evitando falsos positivos.

$$\text{Especificidad} = \frac{\text{VN}}{\text{VN} + \text{FP}}$$

Precisión La precisión (*precision*) señala la proporción de predicciones positivas que efectivamente lo son, aportando información sobre la fiabilidad de las predicciones positivas.

$$\text{Precisión} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

F1-score El F1-score combina la precisión y la sensibilidad en una sola medida (su media armónica), resultando particularmente útil en contextos donde existe un desbalance entre clases.

$$\text{F1-score} = 2 \cdot \frac{\text{Precisión} \cdot \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}}$$

AUC-ROC Por su parte, el AUC-ROC (Área Bajo la Curva *Receiver Operating Characteristic*) permite evaluar la capacidad global del modelo para discriminar entre clases a diferentes umbrales de decisión. La curva ROC se construye graficando la Sensibilidad (eje Y) frente a 1 - Especificidad (eje X). Un valor de AUC cercano a 1.0 indica un rendimiento excelente.

Capítulo 3

Metodologia

3.0.1 Hardware y Software

Todos los procesos, desde el preprocesamiento de los datos y la corrección de los archivos CSV hasta la extracción de características y el entrenamiento de los modelos, se llevaron a cabo en una única estación de trabajo. A continuación, se detallan las especificaciones del hardware y el entorno de software utilizado.

Hardware

Los experimentos se realizaron en un computador de escritorio con una configuración de recursos moderados, cuyas especificaciones son:

- Procesador (CPU): AMD Athlon 3000G
- Memoria RAM: 16 GB DDR4 a 2666 MHz
- Almacenamiento: 2 x Unidades de Estado Sólido (SSD) de 256 GB

Esta configuración de hardware fundamenta algunas de las decisiones metodológicas adoptadas en este estudio, como el uso de imágenes en formato JPG para facilitar el manejo de los datos dentro de un entorno con capacidad computacional limitada.

Software

El análisis de datos y el desarrollo de los modelos se realizaron utilizando el lenguaje de programación Python. El entorno de software se gestionó de la siguiente manera:

- Distribución y Entorno: Se utilizó la distribución Anaconda con el IDE Spyder. Para evitar conflictos de dependencias, se configuraron dos entornos virtuales independientes: uno dedicado exclusivamente a la ejecución de `PyRadiomics` y otro para el resto de los análisis.
- Bibliotecas Principales: Las bibliotecas clave empleadas en este trabajo, junto con su propósito, fueron:
  - Manipulación de Datos: `Pandas` y `NumPy`.
  - Aprendizaje Automático: `Scikit-learn` (sklearn) para la implementación de los modelos SVM y las métricas de evaluación.
  - Extracción de Características Radiómicas: `PyRadiomics`.
  - Aprendizaje Profundo: `TensorFlow` y `Keras` para la extracción de características profundas.
  - Visualización: `Matplotlib` y `Seaborn` para la generación de gráficos y figuras.

Las versiones específicas de cada biblioteca se documentaron para garantizar la total reproducibilidad de los experimentos.

3.0.2 Adquisición y Procesamiento del Conjunto de Datos

Para la adquisición de los datos, se utilizó una versión del conjunto de datos CBIS-DDSM disponible en la plataforma Kaggle. Esta versión se diferencia de la base de datos original, alojada por el NIH (*National Institutes of Health*), principalmente en el formato de las imágenes: las mamografías se encuentran en formato JPG en lugar del estándar médico DICOM.

La elección de este formato representa una decisión metodológica crítica. El formato DICOM es un estándar sin compresión con pérdida que preserva la totalidad de la información radiológica original. En contraste, el formato JPG utiliza un algoritmo de compresión con pérdida (*lossy compression*), lo que implica una reducción del tamaño del archivo (de 163.1 GB a 6.3 GB) a costa de una pérdida irreversible de información en la imagen.

Se reconoce plenamente que esta compresión puede alterar características texturales sutiles que son fundamentales para un análisis radiómico de alta fidelidad. Por lo tanto, el uso de imágenes JPG constituye una limitación fundamental de este estudio, y se asume que las características radiómicas extraídas pueden diferir de las que se obtendrían de los datos DICOM originales.

No obstante, esta decisión fue necesaria para garantizar la viabilidad del estudio dentro de un entorno con recursos computacionales limitados. La versión comprimida, aunque con menor fidelidad, permitió la implementación y evaluación de un rango diverso de modelos de aprendizaje automático. Esta aproximación es consistente con investigaciones recientes que exploran el impacto de la compresión en modelos de clasificación, sugiriendo que, a pesar de la pérdida de información, los patrones diagnósticos a menudo pueden ser aprendidos de manera efectiva [66].

El objetivo, por tanto, es evaluar y comparar el rendimiento de distintos enfoques metodológicos —incluyendo máquinas de vectores de soporte, análisis radiómico y redes neuronales profundas— bajo la restricción impuesta por el uso de datos comprimidos. Este escenario refleja una situación práctica donde el acceso a recursos de alto rendimiento es limitado, aceptando el *trade-off* entre la fidelidad de la imagen y la factibilidad del análisis.

Procesamiento Inicial y Validación del Conjunto de Datos

El conjunto de datos procesado, obtenido de Kaggle, se distribuyó en cuatro archivos CSV que organizaban las lesiones en conjuntos de entrenamiento y prueba, tanto para masas como para calcificaciones. La distribución inicial de las etiquetas en dichos conjuntos se detalla en

la Tabla 3.1. Adicionalmente, se proporcionaron las imágenes de las mamografías completas, las regiones recortadas (*cropped*) y las máscaras de las Regiones de Interés (ROI), tal como se ilustra en la Figura 2.2.

Conjunto	Masas Benig.	Masas Malig.	Calc Benig.	Calc Malig.
Entrenamiento	681	637	1001	544
Prueba	231	147	197	129

Table 3.1: Distribución de lesiones benignas y malignas en los conjuntos de entrenamiento y prueba. Las etiquetas "Benign without call" fueron agrupadas con las benignas.

El siguiente paso fue validar la correspondencia entre los registros de estos archivos CSV y los archivos de imagen (.jpg), ya que una inspección inicial reveló que los CSV contenían rutas de archivo incorrectas y referencias a nombres de archivo del formato DICOM original que no coincidían con los archivos JPG proporcionados. Para resolver esta inconsistencia, se implementó un procedimiento de mapeo y validación sistemático.

El proceso de corrección se realizó de la siguiente manera:

(1) Imágenes de Mamografías Completas: La correspondencia para estas imágenes fue directa, estableciendo un mapeo inequívoco al ser el único archivo `.jpg` en su respectiva carpeta.

(2) Imágenes Recortadas y Máscaras de ROI: Este mapeo presentó una mayor complejidad. Inicialmente, se exploró una heurística basada en el tamaño del archivo, pero la validación demostró que este método era propenso a errores, identificándose una correspondencia incorrecta en 232 casos de calcificaciones.

- Para asegurar la integridad de los análisis preliminares, algunos de los modelos iniciales se desarrollaron utilizando un subconjunto de datos que excluía estos casos ambiguos, registrando debidamente las imágenes omitidas para garantizar la reproducibilidad de dichos experimentos.
- Posteriormente, un análisis más profundo de la nomenclatura reveló el criterio

definitivo: los archivos de las imágenes recortadas siempre comenzaban con el prefijo `1-`. Esta regla permitió establecer un mapeo fiable para la totalidad de los casos.

A modo de ejemplo práctico, el script de procesamiento transformaba una ruta original del CSV, como `.../000000.dcm`, en la ruta local verificada `.../1-213.jpg`, utilizando el identificador numérico del caso para localizar la carpeta y la regla del prefijo `1-` para distinguir entre la imagen recortada y la mascara en el caso de que hubieran dos archivos jpg en una carpeta.

Cabe mencionar que para asegurar una trazabilidad robusta a lo largo de todo el estudio, se generó un identificador único (`caseid`) para cada lesión, vinculando de forma unívoca los registros del CSV con sus respectivas imágenes.

Es importante destacar que los experimentos presentados en este trabajo se realizaron sobre diferentes configuraciones del conjunto de datos para evaluar distintos escenarios. Se entrenaron modelos utilizando únicamente lesiones de masas, únicamente lesiones de calcificaciones, o una combinación de ambas.

Además, debido al proceso de validación de datos descrito previamente, se generaron dos versiones del conjunto de datos de calcificaciones. Las definiciones de estos conjuntos, que se utilizarán a lo largo de este documento, se presentan en la Tabla 3.2.

Término Clave	Descripción
Conjunto Completo	Incluye la totalidad de los casos de calcificaciones, una vez que la correspondencia de todos los archivos fue corregida y validada.
Subconjunto	Versión del conjunto de datos que excluye los 232 casos de calcificaciones cuya correspondencia inicial, basada en heurísticas, fue ambigua.

Table 3.2: Definición de los conjuntos de datos de calcificaciones utilizados.

En resultados y análisis posteriores, se especificará claramente en cada experimento qué configuración de datos fue utilizada (masas, calcificaciones o ambas) y si se empleó el

conjunto completo o el subconjunto de calcificaciones, garantizando así la total transparencia y reproducibilidad de los resultados.

3.0.3 Preprocesamiento

El preprocesamiento se aplicó directamente sobre las imágenes ya recortadas (*cropped*), que contienen exclusivamente la región de interés (ROI). Si bien el recorte elimina eficazmente el tejido mamario circundante y los artefactos externos a la ROI, no aborda el ruido intrínseco presente *dentro* de la propia lesión o en el tejido adyacente inmediato. Este ruido puede manifestarse como píxeles de intensidad anómala (ruido impulsivo o "sal y pimienta"), a menudo como resultado del proceso de adquisición o de la compresión JPG.

Por esta razón, se implementó una secuencia de filtrado post-recorte para mejorar la calidad de la señal dentro de la ROI:

(1) Filtro de Mediana (Median Filter): Se aplicó un filtro de mediana (*kernel* 3x3) para suprimir el ruido impulsivo interno. A diferencia del recorte, que es una operación espacial, el filtro de mediana es una operación de vecindad que homogeneiza la textura local sin comprometer los bordes estructurales de la lesión. Su función aquí es crucial para evitar que píxeles anómalos aislados sean interpretados erróneamente como microcalcificaciones o características radiómicas significativas.

(2) Ecualización Adaptativa (CLAHE): Seguidamente, se aplicó CLAHE (con `clipLimit=2.0` y `tileGridSize=(8,8)`) para realzar el contraste local dentro de la ROI ya filtrada, mejorando la visibilidad de los patrones texturales que definen la lesión.

Este enfoque secuencial, que combina el aislamiento de la ROI, la limpieza de ruido interno y el realce del contraste, es una práctica común que ha demostrado ser efectiva en estudios recientes de análisis de imágenes mamográficas [67, 36, 68], preparando así una imagen de alta calidad para la extracción de características.

3.0.4 Aumentación de Datos del Conjunto de Entrenamiento

Para incrementar la diversidad de los datos de entrenamiento y reducir el sobreajuste (*overfitting*), se implementó un proceso de aumentación de datos.

Para garantizar la validez de los resultados, se siguió una secuencia de operaciones fundamental: primero, el conjunto de datos original fue dividido en los conjuntos de entrenamiento y prueba. A continuación, la aumentación de datos se aplicó exclusivamente sobre las imágenes del conjunto de entrenamiento. Las transformaciones incluyeron rotaciones ($\pm 15°$), traslaciones ($\pm 10\%$), escalamiento ($\pm 15\%$) y ajustes de contraste.

Las nuevas imágenes generadas fueron guardadas en el almacenamiento local, creando así una versión extendida del conjunto de entrenamiento. Para mantener la integridad de este nuevo conjunto, se actualizó el archivo CSV correspondiente, asegurando que cada imagen aumentada heredara el `caseid` y la etiqueta de su imagen original para garantizar la trazabilidad.

Finalmente, y de manera crucial, el conjunto de prueba se mantuvo inalterado durante todo este proceso. De esta forma, se aseguró que la evaluación final del rendimiento de los modelos se realizara sobre datos completamente no vistos, previniendo el sesgo de información (*data leakage*) y obteniendo una medida imparcial de su generalización.

Una vez finalizado el procesamiento y la aumentación de datos, se inició la extracción de características radiómicas y profundas. Descriptores que serán la pieza clave para el entrenamiento de los clasificadores de lesiones mamarias.

3.0.5 Extracción de Características Radiómicas

La extracción de características radiómicas se llevó a cabo utilizando la biblioteca PyRadiomics (versión 3.1.0) en un entorno de Python (versión 3.9.21). Para procesar sistemáticamente todo el conjunto de datos, se adaptó el flujo de trabajo de procesamiento por lotes (*batch processing*) recomendado por los desarrolladores de la librería.

Si bien se utilizó el script `batchprocessing.py` como base, el archivo de configuración

de parámetros, `params.yml`, fue modificado y personalizado para los objetivos específicos de este estudio, en lugar de emplear una configuración por defecto. Los ajustes clave fueron los siguientes:

(1) Número de Bins para Discretización: Se fijó el número de bins en 64 para la discretización de los niveles de gris, un parámetro fundamental que afecta a todas las características de textura. Esta elección se justifica por dos motivos principales:

- Primero, se alinea con la configuración utilizada en estudios previos de radiómica en mamografía, como el de [69], lo que favorece la consistencia y comparabilidad con la literatura existente.
- Segundo, aunque este parámetro es importante, trabajos como el de [31] sugieren que, dentro de un rango razonable (e.g., 32, 64, 128), la elección específica no siempre produce diferencias estadísticamente significativas en el rendimiento final, lo que confiere robustez a la selección de 64 como un valor estándar y aceptado.

(2) Configuración de la Máscara: Se ajustó un parámetro técnico para asegurar la correcta interpretación de las máscaras de la Región de Interés (ROI). Las máscaras procesadas utilizaban el valor 255 (blanco) para delinear la ROI de la imagen recortada, pero la configuración por defecto de PyRadiomics esperaba el valor 1. Inicialmente, se validó el flujo de trabajo convirtiendo las máscaras al formato esperado. Sin embargo, la solución definitiva, utilizada posteriormente, fue modificar el parámetro `label` en la configuración para que aceptara directamente el valor 255, eliminando la necesidad de re-procesar los archivos de imagen.

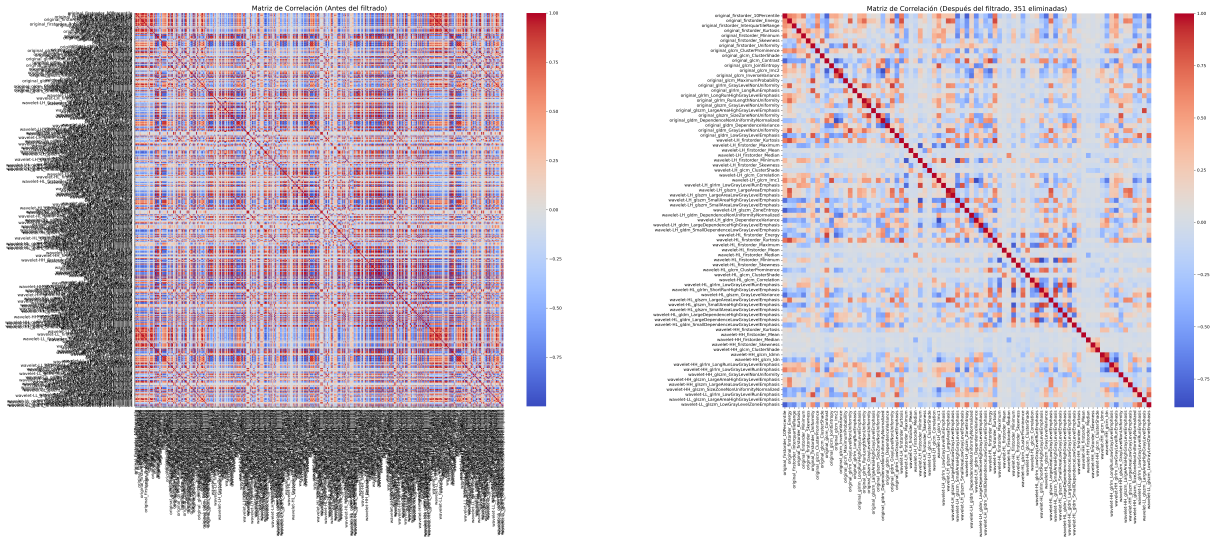
(3) Transformaciones de Imagen: Adicionalmente, se habilitó la aplicación de transformaciones de imagen basadas en filtros *wavelet* para enriquecer el conjunto de características extraídas.

3.0.6 Selección de Características Basada en Correlación

Para reducir la redundancia en el conjunto de 430 características radiómicas extraídas y mejorar la eficiencia de los modelos posteriores, se aplicó un filtro basado en el coeficiente de correlación de Pearson. Se estableció un umbral de $|r| > 0.9$ para identificar y eliminar características con alta colinealidad. Esta técnica asume que si dos variables están fuertemente correlacionadas, contienen información redundante, por lo que una de ellas puede ser eliminada sin una pérdida significativa de poder predictivo, beneficiando la estabilidad del modelo.

El resultado de este procedimiento fue la eliminación de 351 características (81.6%), conservando un conjunto final y más robusto de 79 características [9]. Un análisis de las variables descartadas reveló que la mayoría pertenecían a la familia de transformadas wavelet. Este hallazgo es consistente con la naturaleza de estos filtros, ya que diferentes combinaciones aplicadas a la misma región a menudo capturan patrones texturales muy similares, resultando en una alta correlación.

El efecto de este filtrado es visualmente evidente en la Figura 3.1. La matriz de correlación inicial (a) muestra pequeñas diagonales rojas y otras áreas rojizas fuera de la diagonal principal, indicando una fuerte interdependencia entre múltiples características. Tras el filtrado (b), la matriz se vuelve predominantemente más clara, lo que demuestra que las 79 características restantes poseen una baja correlación lineal entre sí. Nuevamente, para garantizar la trazabilidad y reproducibilidad, se guardaron los archivos de qué características eran eliminadas.



(a) Antes del filtrado (430 características) **(b)** Después del filtrado (79 características)

Figure 3.1: Matriz de correlación de características radiómicas según el coeficiente de Pearson. Se observa una reducción notable de las correlaciones altas tras el proceso de filtrado.

3.0.7 Extracción de Características Profundas mediante Aprendizaje por Transferencia

Para la extracción de características de aprendizaje profundo , se adoptó una estrategia de aprendizaje por transferencia , utilizando dos arquitecturas de Redes Neuronales Convolucionales (CNNs) canónicas: ResNet50 y VGG16.

La elección de VGG16 y ResNet50 no fue arbitraria. Ambas son arquitecturas de referencia en el campo de la visión por computador, pero representan paradigmas de diseño distintos. VGG16 es un modelo secuencial profundo clásico, mientras que ResNet50 introdujo las “conexiones residuales” para permitir el entrenamiento de redes mucho más profundas. Utilizar ambas permite evaluar y comparar la efectividad de dos enfoques arquitectónicos fundamentales.

Se emplearon las versiones de estos modelos pre-entrenadas en el conjunto de datos ImageNet. Aunque ImageNet contiene imágenes naturales y no médicas, el fundamento del aprendizaje por transferencia se basa en la jerarquía de características que estas redes

aprenden. Las capas iniciales de los modelos entrenados en ImageNet aprenden a detectar características genéricas y de bajo nivel (bordes, texturas, colores), las cuales son universales para cualquier tarea de visión, incluidas las imágenes mamográficas. Al reutilizar estos detectores de características ya entrenados, se aprovecha un conocimiento robusto que sería imposible de aprender desde cero con un conjunto de datos médicos, que es inherentemente más pequeño [70, 71].

En este trabajo, se optó por utilizar las redes pre-entrenadas como extractores de características fijos, sin realizar un reentrenamiento o ajuste fino (*fine-tuning*) de sus pesos. Esta es una decisión metodológica deliberada, justificada principalmente por el tamaño del conjunto de datos. Esta estrategia ha sido utilizada en otros estudios, Huynh et al utilizaron AlexNet como extractor de características para entrenar un clasificador híbrido entre radiómica y deep learning[16].

El ajuste fino de una red profunda requiere una cantidad considerable de datos específicos del dominio para evitar el sobreajuste o el "olvido catastrófico" de las características aprendidas. Dado el tamaño limitado de los datos disponibles, congelar los pesos de las capas convolucionales y usarlas únicamente para generar descriptores de las imágenes fue la estrategia más conservadora y robusta. Esto permite aprovechar el poder de las características aprendidas en ImageNet sin correr el riesgo de que el modelo simplemente memorice el conjunto de entrenamiento.

Para condensar los mapas de características generados por las capas convolucionales en un vector de tamaño fijo, se utilizó una capa de `GlobalAveragePooling2D` (GAP). Se prefirió esta técnica sobre otras alternativas por las siguientes razones:

- Frente a `Flatten`: Una capa `Flatten` tradicional desdobla todos los mapas de características en un vector extremadamente largo, lo que aumenta masivamente el número de parámetros y el riesgo de sobreajuste. GAP, en cambio, calcula el promedio de cada mapa de características, generando un vector mucho más compacto y manejable.

- Frente a `GlobalMaxPooling`: Mientras que el *max pooling* solo considera el valor de

activación más alto de cada mapa (útil para detectar características muy localizadas), el *average pooling* proporciona una representación más holística y espacialmente invariante de la presencia de una característica en toda la región de interés.

De esta manera, GAP actúa como un regularizador estructural, reduce drásticamente la dimensionalidad y genera un vector de características robusto, ideal para ser utilizado como entrada en un clasificador posterior como un SVM.

Así,para cada una de estas capas, se entrenó un clasificador binario (SVM) en el subconjunto de datos que combina masas y calcificaciones. El rendimiento de cada capa como extractora de características se evaluó utilizando la métrica del Área Bajo la Curva (AUC). Finalmente, la capa que alcanzó el AUC más alto fue seleccionada como la fuente definitiva de características profundas para los experimentos posteriores. La discusión detallada sobre el rendimiento de cada capa y la justificación de la elección final se presenta en la sección de resultados.

3.0.8 Clasificadores SVM

3.0.9 Modelado y Optimización con SVM

Para la clasificación de las lesiones se construyeron modelos de SVM con la biblioteca `scikit-learn`. El desarrollo se dividió en dos etapas: una fase exploratoria para comprender las características del problema y una fase definitiva para la construcción de los modelos optimizados.

Fase Experimental Exploratoria

La implementación de cada modelo SVM se encapsuló en un Pipeline de la biblioteca scikit-learn. Este consistía en dos pasos secuenciales: primero, un escalado de características mediante StandardScaler para estandarizar los datos a una media de cero y desviación estándar de uno, un paso crucial para el rendimiento de los SVM. A continuación, se definía

el clasificador SVC. Para garantizar la reproducibilidad de todos los experimentos, se fijó un estado aleatorio en el clasificador (`randomstate`=42).

Se llevó a cabo una amplia fase de experimentación preliminar, principalmente sobre el subconjunto de datos. El objetivo era evaluar el impacto de distintas configuraciones metodológicas en el rendimiento de los clasificadores. Esta fase incluyó el entrenamiento y evaluación de múltiples escenarios:

- Modelos binarios para solo masas, solo calcificaciones y para ambos tipos de lesión combinados.
- Un modelo multi-clase (masa benigna/maligna, calcificación benigna/maligna) para investigar la separabilidad entre las cuatro categorías.
- Iteraciones de los modelos anteriores con y sin ciertos pasos de preprocesamiento (como el filtro de mediana y CLAHE) y con y sin aumentación de datos.

Estrategia de Modelado Definitiva

Los aprendizajes obtenidos en la fase exploratoria guiaron la estrategia final. Se concluyó que el enfoque más robusto y comparable con la literatura consistía en desarrollar dos modelos binarios especializados y optimizados: uno para la clasificación de masas y otro para la de calcificaciones. Los resultados principales de este trabajo se centran en estos dos modelos.

Para optimizar la configuración de los modelos, se realizó una búsqueda de hiperparámetros mediante `GridSearchCV` con validación cruzada de 10 pliegues. Considerando el desbalance de clases en el conjunto de calcificaciones, se utilizó el parámetro `classweight='balanced'` y se optimizó la búsqueda en función del F1-score. Sin embargo, para evitar el sobreajuste, este proceso no se aplicó a los modelos que utilizaban características extraídas de redes profundas. Finalmente, para los resultados en el conjunto de prueba, los intervalos de confianza se calcularon mediante la técnica de *bootstrapping* con 1000 iteraciones.

Para garantizar la máxima claridad y reproducibilidad, en la sección de resultados se

especificará de forma explícita la configuración exacta utilizada para cada experimento reportado (e.g., tipo de lesión, conjunto de datos completo o subconjunto, y pasos de preprocesamiento aplicados).

3.0.10 Ensamble

Luego, con el fin de aprovechar la complementariedad entre los modelos entrenados con características radiómicas y profundas, se implementó un esquema de ensamble por soft voting. Para ello, se utilizaron las salidas probabilísticas de dos modelos SVM previamente entrenados —uno sobre características radiómicas y otro sobre características profundas extraídas con redes convolucionales preentrenadas—. Se evaluaron combinaciones convexas de estas salidas, buscando los pesos óptimos que maximizaban el AUC macro mediante una búsqueda exhaustiva sobre el espacio de pesos $(w_{\text{rad}}, w_{\text{deep}})$, discretizado en pasos de 0.1. Una vez hallada la combinación con mejor rendimiento, se calcularon las predicciones finales como la clase con mayor probabilidad en el vector resultante.

3.0.11 Stacking

Para integrar la información proveniente de las características radiómicas y las características profundas, se implementó una estrategia de stacking de dos niveles. En el primer nivel se emplearon dos clasificadores SVM previamente optimizados: uno entrenado con las características radiómicas extraídas de las mamografías mediante PyRadiomics y otro entrenado con los vectores de características profundas obtenidos a partir de ResNet-50. Con el fin de evitar sobreajuste y fuga de información, se utilizó validación cruzada estratificada de 10 pliegues, generando predicciones out-of-fold (OOF) de cada modelo base, que posteriormente fueron empleadas como variables de entrada para el modelo de segundo nivel. En esta etapa, se entrenó una regresión logística binaria, cuyo objetivo fue aprender a combinar las salidas de ambos clasificadores base para mejorar la capacidad predictiva del sistema. Finalmente, ambos SVM fueron reentrenados con el conjunto completo de entrenamiento

y se construyó un clasificador apilado final que encapsula tanto los modelos base como el meta-modelo, permitiendo realizar predicciones sobre nuevos casos de forma equivalente a un único clasificador.

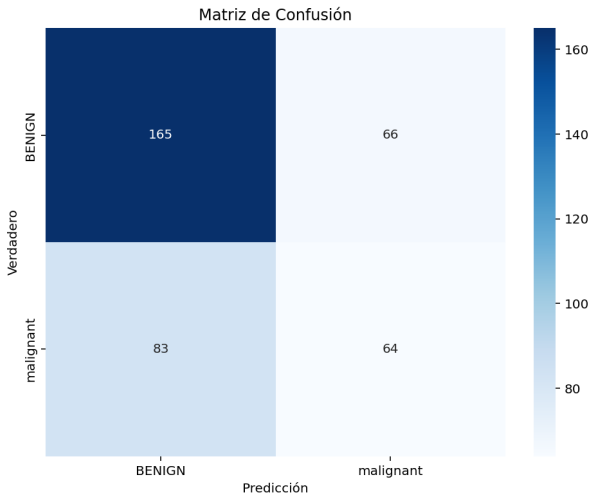
Capítulo 4

Resultados y análisis

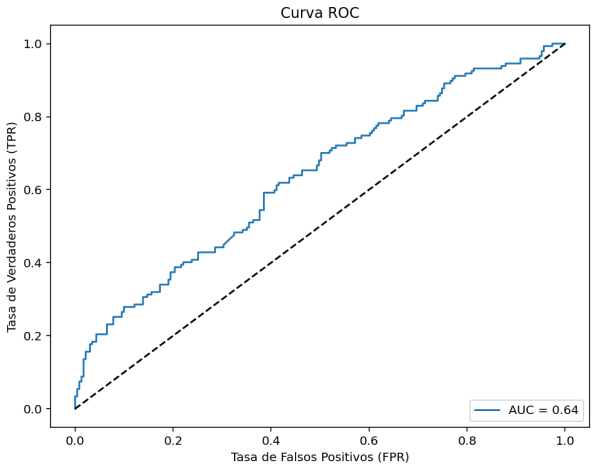
Algunos experimentos preliminares se realizaron con el objetivo de evaluar si el preprocesamiento mejoraba el desempeño del modelo. En la Figura 4.1 se presenta la matriz de confusión y la curva ROC correspondientes a un modelo SVM entrenado utilizando únicamente las características de textura extraídas del *subconjunto* de imágenes de masas del CBIS-DDSM (vease la tabla 3.2). Es importante aclarar que estos resultados se obtuvieron sin aplicar varias etapas clave del pipeline, tales como: el ajuste a 64 bins para la cuantificación de intensidades, la extracción de características radiómicas basadas en wavelets, la selección de características, la validación cruzada (cross-validation) y la búsqueda de hiperparámetros mediante grid search. Por tanto, estos resultados deben interpretarse únicamente como una línea base inicial. Asi, de este primero modelo, se reportó un rendimiento de la curva AUC de 0.64, y en la matriz de confusión se puede observar una predisposicion a clasificar las lesiones como benignas.

Siguiendo la misma línea de trabajo, se entrenó un modelo SVM utilizando el subconjunto de imágenes de masas, esta vez a partir de las características profundas extraídas de la última capa de una red ResNet-50 preentrenada en ImageNet. Los resultados se resumen en la Tabla 4.1.

Continuando con el análisis sobre el subconjunto de imágenes de masas, se evaluó el



(a) Modelo SVM sin preprocesamiento ni datos adicionales.



(b) Curva AUC correspondiente al modelo.

Figure 4.1: Resultados del modelo SVM con características radiómicas sin preprocesamiento ni datos adicionales entrenados en el subconjunto de masas.

Clase	**Precisión**	**Recall**	**F1-score**	**Soporte**
0	0.75	0.73	0.74	231
1	0.59	0.62	0.60	147
Exactitud			**0.69**	378
Promedio macro	0.67	0.67	0.67	378
Promedio ponderado	0.69	0.69	0.69	378

Table 4.1: Métricas de desempeño del clasificador basado en la ultima capa de resnet 50 preentrenada en el subconjunto de prueba de las masas. Las clases 0 y 1 corresponden a benigno y maligno respectivamente. **AUC** = 0.7468.

desempeño del modelo SVM empleando características profundas extraídas de diferentes capas de la red VGG-16 preentrenada en ImageNet. Para ello, se generaron vectores de características a partir de cada capa convolucional de la red, con el fin de identificar cuál de ellas ofrecía una mejor representación para la tarea de clasificación. Se encontró que una capa intermedia, específicamente la segunda convolución del bloque 4 (block4conv2), arrojó el mejor desempeño, con un valor de AUC cercano a 0.75.

De manera análoga, se repitió el procedimiento utilizando una red ResNet-50 preentrenada, esta vez sobre el subconjunto completo que incluye tanto masas como calcificaciones, aplicando

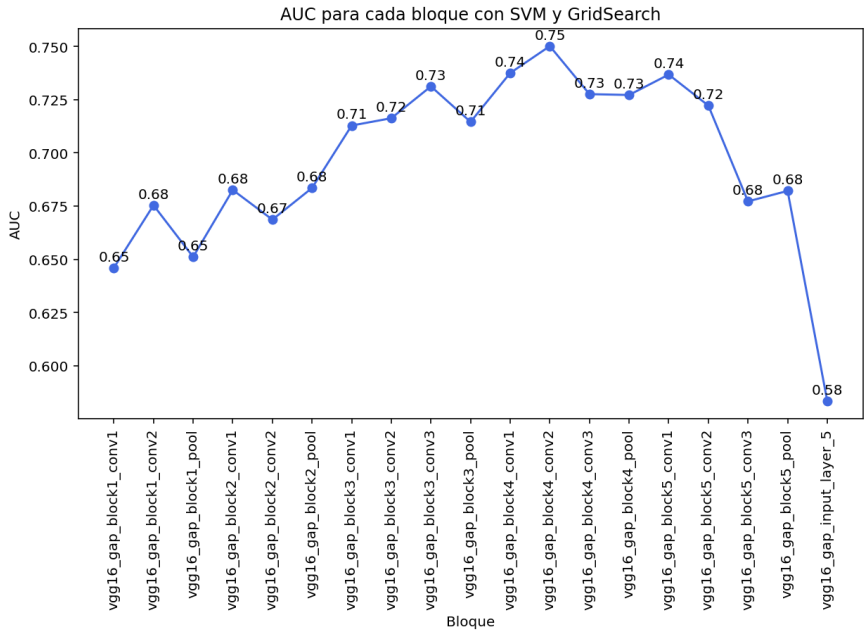


Figure 4.2: Capas de la red VGG-16 vs clasificaciones de masas malignas o benignas (AUC como metrica de rendimiento.)

el preprocesamiento correspondiente y técnicas de data augmentation. Al igual que en el caso de VGG-16, se exploraron distintas capas de la red para la extracción de características profundas. Nuevamente se observó que la última capa no necesariamente proporciona la mejor representación para la tarea de clasificación. En este caso, una capa intermedia resultó en el mejor desempeño, alcanzando un valor de AUC de 0.73.

Los resultados de las figuras 4.3 y 4.2 demuestran que las características de las capas intermedias de VGG-16 y ResNet-50 proporcionan un rendimiento superior para la clasificación con SVM en comparación con las capas finales. Este fenómeno subraya un aspecto clave del aprendizaje por transferencia: la relevancia de seleccionar la profundidad adecuada del extractor de características. Dicha observación concuerda con los hallazgos de estudios anteriores, los cuales, al aplicar técnicas similares, también concluyeron que las capas internas de una red preentrenada pueden generar representaciones más ricas y discriminativas para la clasificación de lesiones mamarias [16].

Otro modelo de clasificación basado en características radiómicas (sin aplicación de selección de características) fue entrenado sobre el subconjunto de masas. En este caso, se

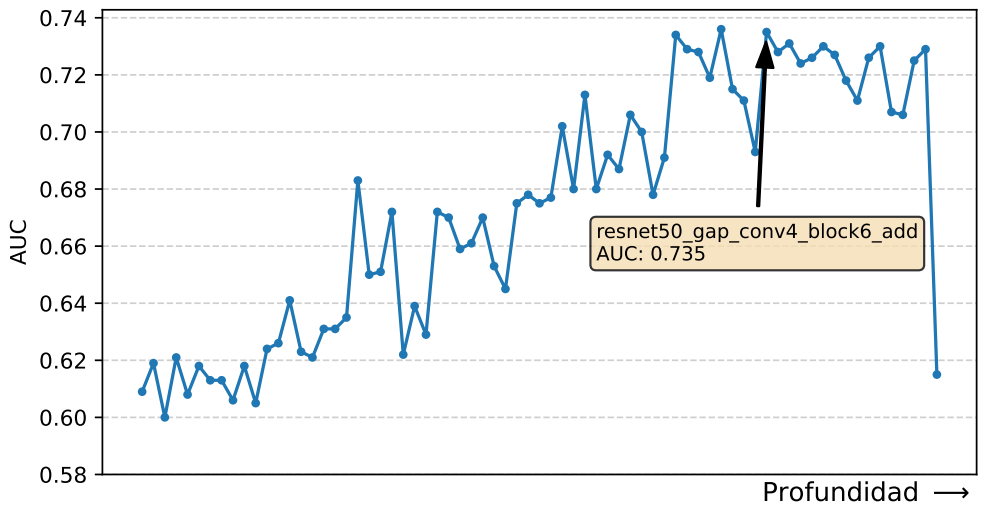
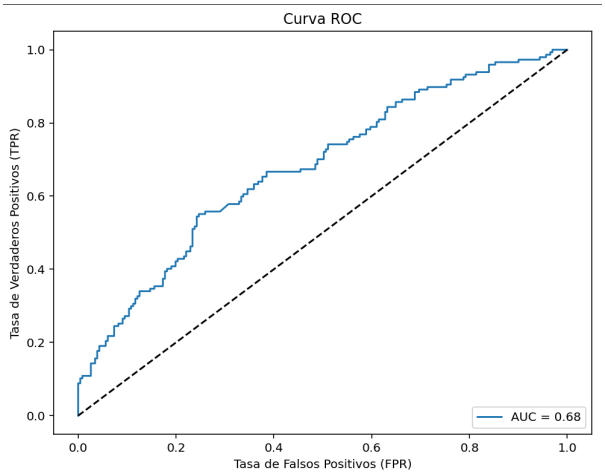


Figure 4.3: AUC de los SVM entrenados a partir de las capas internas de una red resnet 50 preentrenada en Imagenet utilizando el subconjunto de masas y calcificaciones.

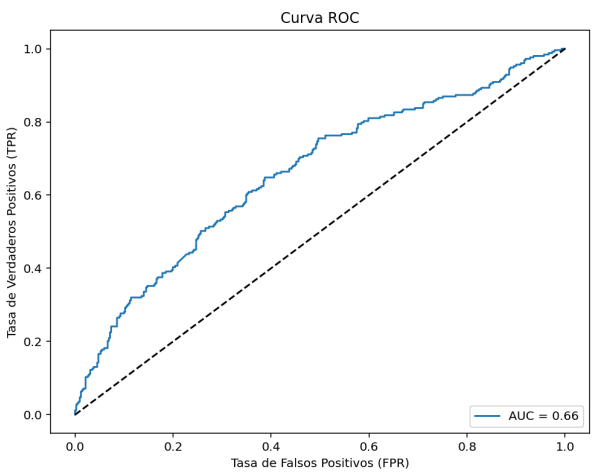
obtuvo un AUC de 0.68, lo que representa una mejora respecto a los resultados mostrados en la Figura 4.1b, atribuible al preprocesamiento aplicado a las imágenes. No obstante, como se observa en la Figura 4.4b, el rendimiento del modelo disminuye al incluir también imágenes de calcificaciones. Esta caída en el desempeño puede deberse al aumento de la heterogeneidad del conjunto de datos, entendida como la mayor diversidad en los patrones de intensidad, textura y morfología entre los distintos tipos de lesiones. Mientras que las masas suelen presentar bordes menos definidos y estructuras más difusas, las calcificaciones tienden a ser más pequeñas, de alta densidad y con formas distintas, lo que puede dificultar que un mismo conjunto de características radiómicas capture eficazmente la información relevante para ambas clases.

Dado que la inclusión de calcificaciones en los modelos supervisados afectó negativamente el rendimiento, se optó por un enfoque no supervisado para explorar la estructura subyacente del conjunto de datos combinado. Con este fin, se entrenó un modelo *k-means* utilizando las características radiómicas extraídas tanto de masas como de calcificaciones.

La Figura 4.5b presenta el resultado del método del codo, empleado para estimar el número óptimo de clústeres. La gráfica evidencia que una partición en dos grupos ($k = 2$),



(a) Modelo SVM entrenado con características radiómicas en masas preprocesadas.



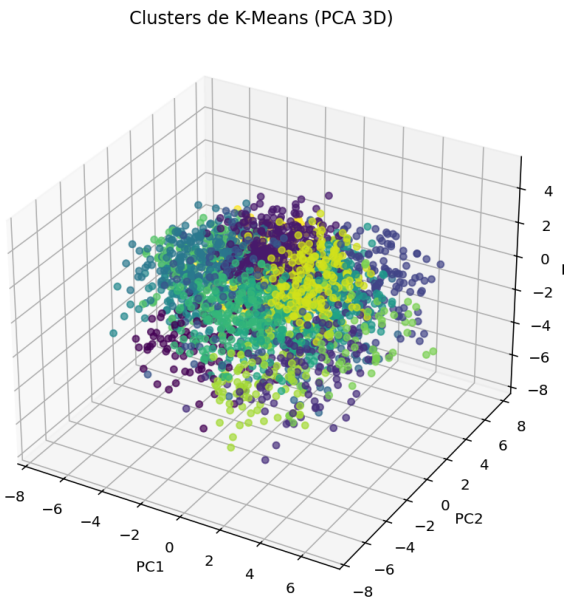
(b) Modelo SVM entrenado con características radiómicas en masas y calcificaciones preprocesadas.

Figure 4.4: Comparación del desempeño del SVM con distintos subconjuntos de entrenamiento preprocesados.

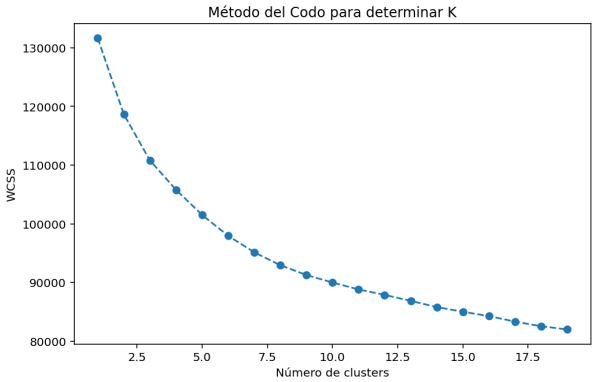
equivalente a una clasificación binaria, es insuficiente para capturar la complejidad de los datos. La técnica del codo sugiere que el número apropiado de clústeres se encuentra donde un incremento en k genera una disminución marginal en la inercia (WCSS), lo cual ocurre aproximadamente en $k = 16$. Esta granularidad tan alta se explica porque las masas se separan principalmente por su forma (e.g., irregular, ovalada) y márgenes (e.g., espiculados, circunscritos), mientras que las calcificaciones se diferencian por su morfología (e.g., finas lineales, pleomórficas) y distribución (e.g., agrupada, segmental).

Sin embargo, proponer un modelo multimodal con un número tan elevado de categorías es inviable debido a las limitaciones en el tamaño del conjunto de datos. En su lugar, se consideró que una división en cuatro clústeres ($k = 4$) representa un equilibrio más adecuado. Esta elección se fundamenta en la hipótesis de que dichos grupos podrían corresponderse con las cuatro clases de interés clínico: masas malignas, masas benignas, calcificaciones malignas y calcificaciones benignas, permitiendo así una interpretación clínicamente relevante de los subgrupos identificados.

Se repitió el proceso utilizando el subconjunto de imágenes que incluye masas y calcificaciones. A partir de este conjunto, se entrenaron modelos SVM basados en características



(a) Resultados del agrupamiento mediante K-means.



(b) Método del codo para determinar el número óptimo de clusters.

Figure 4.5: Agrupamiento no supervisado usando K-means y análisis del número de clusters.

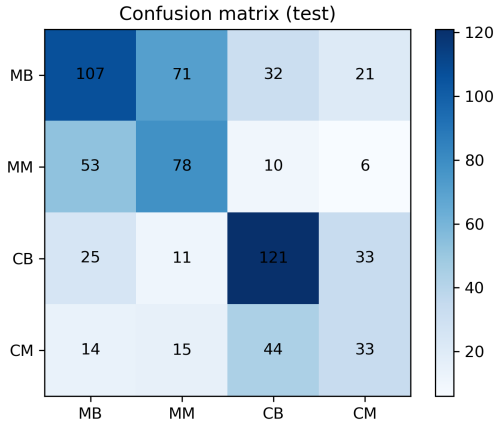
radiómicas, modelos basados en deep learning y un modelo de ensamble mediante soft voting, siguiendo el mismo protocolo establecido previamente, con la única diferencia de no aplicar técnicas de data augmentation ni gridsearch. Los resultados obtenidos se resumen en la Tabla 4.2. En ella se evidencia que el rendimiento general de los modelos se ve considerablemente afectado por la inclusión de las calcificaciones, siendo especialmente crítica la clasificación de las calcificaciones malignas, cuyo F1-score se encuentra alrededor de 0.3. Este comportamiento refuerza la idea de que las calcificaciones representan un reto adicional para los clasificadores, posiblemente debido a su menor representación en el conjunto de datos, así como a su morfología más sutil y difícil de caracterizar mediante los descriptores utilizados.

Una posible solución a las limitaciones observadas en la clasificación de calcificaciones malignas sería la incorporación de características radiómicas de forma al modelo basado en radiomica. Sin embargo, esto requeriría modificar el enfoque actual de extracción de características. La biblioteca PyRadiomics, empleada en este trabajo, necesita tanto la imagen original como la máscara correspondiente para calcular adecuadamente las características de forma, ya que muchas de estas dependen no solo de los valores de píxeles dentro del

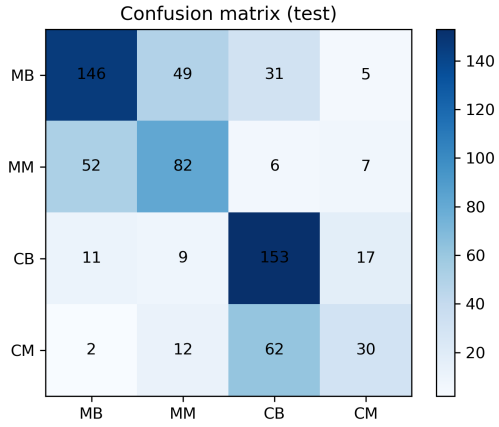
Categoría	Métrica	SVM Radiómica	SVM DL	Soft voting
masa benigna	F1-score	0.50	0.66	0.67
	AUC	0.7467	0.8458	—
masa maligna	F1-score	0.48	0.55	0.58
	AUC	0.7970	0.8433	—
calc. benigna	F1-score	0.61	0.69	0.69
	AUC	0.8432	0.8847	—
calc. maligna	F1-score	0.33	0.36	0.32
	AUC	0.7483	0.8241	—
macro avg	F1-score	0.48	0.57	0.57
	AUC	0.7838	0.8495	0.8594
weighted avg	F1-score	0.50	0.60	0.60
	AUC	0.7945	0.8532	—

Table 4.2: Comparación de desempeño por clase y promedios entre modelos (subconjunto de prueba, sin aumento de datos y sin optimización de hiperparametros)

ROI, sino también de su contexto en la imagen circundante. En el enfoque actual, donde se utilizan imágenes recortadas (cropped) centradas en el ROI, la máscara cubre la totalidad de la imagen, lo cual impide calcular la relación espacial con el tejido adyacente. Por tanto, sería necesario volver a trabajar sobre las imágenes completas y sus máscaras originales para habilitar este tipo de análisis. Cabe destacar que algunas características morfológicas, como el perímetro, han demostrado ser relevantes en la identificación de calcificaciones, como se evidencia en el estudio de Stefano et al. [68].



(a) Matriz de confusión del modelo SVM utilizando características radiómicas.



(b) Matriz de confusión del modelo SVM utilizando características profundas (deep features).

Figure 4.6: Comparación de matrices de confusión del clasificador SVM en la tarea de clasificación de 4 clases.

La observación de las matrices de confusión para el modelo de cuatro clases, presentadas en la figura 4.6, revela la presencia de una cantidad significativa de elementos fuera de la diagonal principal. Este patrón sugiere que el modelo confunde las clases, un problema que podría mitigarse desarrollando dos clasificadores especializados: uno para masas y otro para calcificaciones. Aunque esta estrategia implica sacrificar la generalidad del modelo original de cuatro clases, que no requiere una distinción previa entre los tipos de hallazgo, ofrece la posibilidad de mejorar el rendimiento diagnóstico de manera considerable.

Siguiendo esta línea, se entrenaron nuevamente los modelos SVM, tanto radiómicos como profundos, utilizando esta vez el conjunto completo de datos para cada tarea específica. Los resultados obtenidos se detallan en las tablas 4.3 y 4.4. Adicionalmente, se exploró el uso de un modelo de ensamblaje mediante stacking para el conjunto de masas. Este enfoque logró una leve mejora, elevando el AUC de 0.762 (obtenido con el SVM-ResNet50) a 0.769. Sin embargo, un método de ensamblaje más sencillo como el soft voting demostró ser más efectivo, alcanzando un AUC de 0.774.

Métrica	SVM radiómica		SVM ResNet50		Soft Voting		Stacking	
	Benignas	Malignas	Benignas	Malignas	Benignas	Malignas	Benignas	Malignas
Precisión	0.67	0.50	0.71	0.62	0.71	0.53	0.7228	0.5319
Recall	0.69	0.48	0.80	0.50	0.78	0.44	0.7684	0.4717
F1-score	0.68	0.49	0.75	0.55	0.75	0.48	0.7449	0.5000
AUC (bootstraping)	0.63 (IC95%: 0.56–0.69)		0.74 (IC95%: 0.68–0.79)		0.72		0.71	
F1 (bootstraping)	0.49 (IC95%: 0.41–0.56)		0.65 (IC95%: 0.60–0.70)		0.61		0.50	
Accuracy (bootstraping)	0.60 (IC95%: 0.56–0.66)		0.68 (IC95%: 0.63–0.72)		0.66		0.66	

Table 4.3: Resultados de clasificación para calcificaciones en el conjunto completo de prueba.

Métrica	SVM radiómica		SVM ResNet50		Soft Voting		Stacking	
	Benignas	Malignas	Benignas	Malignas	Benignas	Malignas	Benignas	Malignas
Precisión	0.72	0.52	0.76	0.62	0.77	0.63	0.7521	0.6181
Recall	0.63	0.63	0.76	0.62	0.76	0.65	0.7619	0.6054
F1-score	0.67	0.57	0.76	0.62	0.76	0.64	0.7570	0.6117
AUC (bootstraping)	0.66 (IC95% 0.61-0.71)		0.762 (IC95%: 0.715–0.809)		0.774		0.7693	
F1 (bootstraping)	0.62 (IC95% 0.57-0.67)		0.689 (IC95%: 0.643–0.731)		0.70		0.6843	
Accuracy (bootstraping)	0.62 (IC95% 0.58-0.68)		0.762 (IC95%: 0.715–0.809)		0.71		0.7011	

Table 4.4: Resultados de clasificación para masas en el conjunto completo de prueba.

En marcado contraste con el comportamiento observado para las masas, la aplicación de métodos de ensamble resultó contraproducente para la clasificación de calcificaciones. El rendimiento del mejor modelo individual, el SVM basado en características de ResNet-50 (AUC = 0.74), se vio reducido al aplicar *stacking* (AUC = 0.72) y *soft voting* (AUC = 0.71). Este resultado sugiere que la fusión de modelos introdujo una fuente de error, posiblemente proveniente del clasificador SVM basado en radiómica. El análisis de la Tabla 4.3 respalda esta idea, ya que las métricas de precisión, recall y F1-score de este modelo para las calcificaciones malignas son inferiores a 0.5. Un rendimiento tan bajo, deficiente incluso frente a un clasificador aleatorio, probablemente aportó predicciones erróneas que penalizaron el desempeño global del ensamble.

Este resultado subraya la dificultad intrínseca de clasificar este tipo de lesión. El problema fundamental parece radicar en el pronunciado desbalance de clases, donde la categoría de calcificaciones malignas está subrepresentada. Esta subclase consistentemente produjo métricas de rendimiento pobres, dificultando una clasificación apropiada. Cabe mencionar que esta dificultad persistió a pesar de las estrategias implementadas para mitigar el desbalance, tales como la validación cruzada de 10 pliegues y la optimización del clasificador SVM con el

hiperparámetro `class_weight` durante la búsqueda en rejilla.

Al contrastar los resultados con la literatura que emplea la base de datos CBIS-DDSM, se observa un rendimiento competitivo en la clasificación de masas. Específicamente, como se detalla en la Tabla 4.4, el modelo SVM con características de aprendizaje profundo (SVM+DL) alcanzó un AUC de 0.76, igualando el resultado reportado por Stefanno et al., mientras que el SVM basado en radiómica (AUC de 0.66) superó al de Lauciello et al. Adicionalmente, el modelo híbrido de *soft voting* mejoró el rendimiento en un 1.5% respecto al mejor modelo individual; sin embargo, esta ganancia no representa una superioridad estadísticamente significativa.

Por otro lado, se evidencia una brecha de rendimiento en la clasificación de calcificaciones. Los resultados obtenidos en este trabajo alcanzan aproximadamente el 90.7% y el 84.7% del AUC reportado por Stefanno et al. para sus enfoques de aprendizaje profundo y radiómica, respectivamente.

Estudio	**Lesión**	**Modelo**	**AUC (%)**	**Precisión (%)**
Stefanno et al	Masas	SVM (Radiómica)	66.70	60.00
	Calc.	SVM (Radiómica)	74.31	70.00
	Masas	EfficientNetB6 (DL)	76.24	-
	Calc.	EfficientNetB6 (DL)	81.52	-
Lauciello et al	Calc.	LDA/SVM (Radiómica)	68.28	-
	Masas	LDA/SVM (Radiómica)	61.53	-
Garro	Masas	Ensamble (Soft Voting)	**77.40**	**71.00**
	Calc.	SVM (ResNet50)	**74.00**	**68.00**

Table 4.5: Resultados comparativos entre enfoques de deep learning. Stefanno et al. aplicaron fine-tuning, mientras que este trabajo demuestra la viabilidad del transfer learning como extractor de características, obteniendo resultados competitivos en la tarea de clasificación de masas [68, 72].

4.1 Conclusiones y perspectivas

Este estudio permitió analizar el problema de clasificación de lesiones mamarias, abordándolo tanto desde los métodos convencionales del análisis de imágenes computacional, como es la radiómica, así como desde las tendencias actuales basadas en el aprendizaje profundo.

Con los modelos SVM basados en características radiómicas se obtuvo un AUC de 0.66 (IC 95%: 0.61-0.71) para masas y de 0.63 (IC 95%: 0.56-0.69) para calcificaciones. Este rendimiento es, en general, consistente con la literatura que utiliza la base de datos CBIS-DDSM. Sin embargo, para las calcificaciones se evidencia una brecha de rendimiento, alcanzando aproximadamente el 84.7% del AUC reportado por Stefanno et al. con un enfoque radiómico similar. Esta dificultad se atribuye a dos factores: el desbalance de clases y la pérdida de información de alta frecuencia por la compresión del formato JPG, que afecta a las texturas finas cruciales para las microcalcificaciones.

El modelo SVM que utilizó características de ResNet-50 alcanzó un AUC de 0.76 en la clasificación de masas. En contraste, para las calcificaciones que alcanzaron un rendimiento del 90.7% en comparación con la literatura, confirmando la mayor complejidad de esta clase. Estos resultados validan la efectividad del aprendizaje por transferencia para esta aplicación clínica y, a su vez, respaldan la hipótesis de que las capas intermedias de una red neuronal convolucional pueden ser más idóneas que las finales para la extracción de características.

El enfoque híbrido para la clasificación de masas, mediante un ensamble de *soft voting*, alcanzó el rendimiento más alto con un AUC de 0.774. Este resultado, superior en un 1.5% al mejor modelo de la literatura comparada, demuestra que la información complementaria de las características radiómicas es valiosa. Aunque el aprendizaje profundo es superior individualmente, la combinación de ambos enfoques genera una descripción más rica de las imágenes y, por tanto, una clasificación más precisa.

En este estudio se concluye que los enfoques híbridos son una estrategia efectiva y viable para la clasificación de lesiones mamarias. Para abordar las dificultades de estandarización y reproducibilidad en la radiómica, se fomentó la transparencia compartiendo el código y

los datos procesados en un repositorio público. Además, la metodología desarrollada, al estar adaptada a un entorno de recursos limitados, es extrapolable a bases de datos privadas o puede ser implementada en centros de diagnóstico en regiones rurales con condiciones tecnológicas similares.

Como trabajo futuro, se proponen varias líneas prioritarias para superar las limitaciones identificadas. Primero, se reentrenará con datos DICOM originales para cuantificar el impacto real de la compresión JPG. Segundo, se implementarán técnicas de *few-shot learning* para la clase minoritaria de calcificaciones malignas, que presenta datos escasos (n<150). Tercero, se explorarán arquitecturas híbridas CNN-Transformer, ya que se consideran más adecuadas que los *Vision Transformers* puros para preservar la información espacial local que es crítica para las microcalcificaciones. Finalmente, se desarrollarán métricas de evaluación que ponderen apropiadamente el costo clínico de los falsos negativos frente a los falsos positivos.

Apéndice A

A.1 Construcción de características de segundo orden

A.1.1 GLCM

A partir de la matriz $P(i,j \mid \delta,\theta)$ se pueden calcular las distribuciones de probabilidad de los elementos de la GLCM, en las cuales se basan las características GLCM. Para ello, primero es necesario introducir las siguientes variables, sea:

- ϵ es un número positivo arbitrariamente pequeño ($\approx 2.2 \times 10^{-16}$).
- $\mathbf{P}(i,j)$ es la matriz de co-ocurrencia para un desplazamiento δ y ángulo θ arbitrarios.
- $p(i,j)$ es la matriz de co-ocurrencia normalizada, definida como $p(i,j) = \frac{\mathbf{P}(i,j)}{\sum \mathbf{P}(i,j)}$.
- N_g es el número de niveles de intensidad discretos en la imagen.
- $p_x(i) = \sum_{j=1}^{N_g} p(i,j)$ son las probabilidades marginales de fila.
- $p_y(j) = \sum_{i=1}^{N_g} p(i,j)$ son las probabilidades marginales de columna.
- μ_x es la intensidad media del nivel de gris de p_x, definida como $\mu_x = \sum_{i=1}^{N_g} p_x(i) \cdot i$.
- μ_y es la intensidad media del nivel de gris de p_y, definida como $\mu_y = \sum_{j=1}^{N_g} p_y(j) \cdot j$.

Las aplicaciones más recientes usan hasta 22 características GLCM, Por defecto, en pyradiomics el valor de una característica se calcula en la GLCM para cada ángulo por

separado, tras lo cual se devuelve la media de estos valores, de las cuales las más utilizadas son [38] :

$$\text{Correlación} = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\cdot i\cdot j - \mu_x\mu_y}{\sigma_x\sigma_y} \tag{A.1}$$

La correlación es un valor entre -1 y 1 que muestra la dependencia lineal de los niveles de gris respecto a sus vóxeles correspondientes en la GLCM. Un valor de 1 indica correlación perfecta, mientras que 0 implica no correlación. Las variables μ_x, μ_y son las medias y σ_x, σ_y las desviaciones estándar de las distribuciones marginales. En el contexto clínico, una alta correlación puede indicar patrones organizados en la textura, lo cual es característico en lesiones benignas, mientras que valores bajos reflejan mayor heterogeneidad típica de tumores malignos [44, 73].

$$\text{Contraste} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-j)^2\, p(i,j) \tag{A.2}$$

El contraste mide la variación local de la intensidad, favoreciendo valores alejados de la diagonal ($i = j$). Un valor alto indica mayor disparidad de intensidad entre vóxeles vecinos. Clínicamente, un alto contraste suele asociarse a bordes irregulares y microcalcificaciones, hallazgos que tienden a estar presentes en lesiones malignas.

$$\text{Energía conjunta} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\left(p(i,j)\right)^2 \tag{A.3}$$

La Energía conjunta mide patrones homogéneos en la imagen. Un valor mayor implica que hay más pares de intensidad (i, j) que aparecen como vecinos con alta frecuencia. Desde el punto de vista clínico, valores elevados de energía indican uniformidad en la textura, común en lesiones benignas bien delimitadas, mientras que valores bajos reflejan desorganización estructural propia de tumores malignos [74].

$$\text{Entropía conjunta} = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)\log_2\left(p(i,j)+\epsilon\right) \tag{A.4}$$

La Entropía conjunta mide la aleatoriedad o variabilidad en los valores de intensidad vecinos. ϵ es un valor pequeño ($\approx 2.2\times10^{-16}$) para evitar $\log(0)$. En aplicaciones clínicas, una entropía elevada refleja mayor heterogeneidad en la textura, lo cual es típico en lesiones malignas con arquitecturas internas complejas, mientras que lesiones benignas tienden a mostrar valores menores por su homogeneidad [75].

A.1.2 GLRLM

Para calcular las características GLRLM, primero se calculan en las matrices para cada dirección definida por el ángulo θ y posteriormente se promedian los resultados.

Luego, se definen las siguientes magnitudes: sea N_g el número de niveles de intensidad discretos en la imagen, N_r el número máximo de longitudes de carrera discretas y N_p el número total de vóxeles en la imagen.

El número de carreras en la imagen para una dirección θ se denota como $N_r(\theta)$ y está dado por

$$N_r(\theta) = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r} P(i,j|\theta),$$

cumpliéndose que $1 \leq N_r(\theta) \leq N_p$.

La matriz de longitudes de carrera para una dirección arbitraria θ se denota $P(i,j|\theta)$, mientras que su versión normalizada se expresa como $p(i,j|\theta)$, definida mediante la relación

$$p(i,j|\theta) = \frac{P(i,j|\theta)}{N_r(\theta)}.$$

En base a lo anterior, el equipo de investigación de Galloway propuso un conjunto de cinco características extraídas de la *GLRLM*. Estas métricas, diseñadas para cuantificar las propiedades texturales de una imagen, se detallan a continuación.

- Énfasis en Rachas Cortas (SRE) Esta característica mide la distribución de las rachas

de longitud corta. Un valor elevado de SRE sugiere que en la imagen predominan las texturas finas y los detalles pequeños, propio de lesiones benignas. Su cálculo se realiza mediante la siguiente fórmula:

$$SRE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{P(i,j|\theta)}{j^2}}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} P(i,j|\theta)}$$

- Énfasis en Rachas Largas (LRE): De forma opuesta al SRE, esta métrica cuantifica la prevalencia de rachas de longitud larga. Un valor alto de LRE es indicativo de texturas más homogéneas y estructuralmente gruesas, por lo que entre mayor sea el valor de esta características mas problema es que se trate de una lesión benigna. Se define como:

$$LRE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} P(i,j|\theta) j^2}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} P(i,j|\theta)}$$

- No Uniformidad de Niveles de Gris (GLN) : Esta característica evalúa la similitud en los niveles de intensidad de gris a lo largo de la imagen. Un valor bajo de GLN denota una distribución más homogénea de los niveles de gris en las distintas rachas, en otras palabras entre mayor sea este valor mayor sera la heterogeneidad, lo cual se asocia a lesiones malignas. La ecuación es:

$$GLN = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_r} P(i,j|\theta) \right)^2}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} P(i,j|\theta)}$$

- No Uniformidad de Longitud de Racha (RLN) : Esta métrica mide la variabilidad en las longitudes de las rachas presentes en la imagen. Si el valor de RLN es bajo, significa que las longitudes de las rachas son más uniformes en toda la imagen. Se calcula así:

$$RLN = \frac{\sum_{j=1}^{N_r} \left(\sum_{i=1}^{N_g} P(i,j|\theta) \right)^2}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} P(i,j|\theta)}$$

- Porcentaje de Rachas (RP) : El RP indica la frecuencia o densidad de las rachas en

relación con el número total de píxeles posibles. Un valor alto de RP puede indicar una imagen con una alta fragmentación textural. Un RP elevado puede sugerir que el tejido está compuesto por múltiples patrones de corta extensión, lo cual se asocia con una mayor irregularidad y, en muchos casos, con procesos malignos. Por el contrario, un RP bajo refleja texturas más continuas y extensas, típicas de tejidos mamarios sanos o lesiones benignas bien delimitadas Se define por la relación:

$$RP = \frac{1}{N_p} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} P(i,j|\theta)$$

A.1.3 GLSZM

Para el cálculo de las características GLSZM, es necesario definir primero algunas magnitudes fundamentales. Sea N_g el número de niveles de intensidad discretos en la imagen, N_s el número de diferentes tamaños de zona presentes, y N_p el número total de vóxeles en la región de interés (ROI). La Matriz de Zonas por Tamaño y Nivel de Gris se denota como $P(i,j)$, donde cada elemento (i,j) representa el número de zonas con un nivel de intensidad i y un tamaño de j vóxeles. A partir de esta matriz, se calcula el número total de zonas, N_z, mediante la suma de todos sus elementos:

$$N_z = \sum_{i=1}^{N_g} \sum_{j=1}^{N_s} P(i,j),$$

cumpliéndose la condición de que $1 \leq N_z \leq N_p$. Finalmente, para obtener la matriz normalizada $p(i,j)$, cada elemento de la matriz original se divide por el número total de zonas N_z, según la relación:

$$p(i,j) = \frac{P(i,j)}{N_z}.$$

Las formulaciones matemáticas de las características GLSZM son análogas a las definidas previamente para GLRLM.

Bibliografía

[1] P. A. Solano-Dazzarola, G. A. Grilló, J. A. López, and E. Montoya-Cobo. "Panorama colombiano del cáncer de mama, cérvix y próstata". *Salutem Scientia Spiritus* 9 (May 2023), 28–35.

[2] World Health Organization. *WHO Position Paper on Mammography Screening*. Geneva: World Health Organization, 2014.

[3] *Radiologist Shortage Will Persist into 2055 without Counteraction*. Feb. 2025.

[4] P. R. O. Tellez, D. A. R. Miranda, and D. A. C. Ortiz. "Estimación de oferta de médicos especialistas en Colombia 1950-2030 Anexo metodológico" ().

[5] N. Eisemann et al. "Nationwide Real-World Implementation of AI for Cancer Detection in Population-Based Mammography Screening". *Nature Medicine* 31 (Mar. 2025), 917–924.

[6] R. J. Gillies, P. E. Kinahan, and H. Hricak. "Radiomics: Images Are More than Pictures, They Are Data". *Radiology* 278 (Feb. 2016), 563–577.

[7] *Artificial Intelligence - Worldwide | Market Forecast*.

[8] H. Li, K. R. Mendel, L. Lan, D. Sheth, and M. L. Giger. "Digital Mammography in Breast Cancer: Additive Value of Radiomics of Breast Parenchyma". *Radiology* 291 (Apr. 2019), 15–20.

[9] N. Mao, P. Yin, Q. Wang, M. Liu, J. Dong, X. Zhang, H. Xie, and N. Hong. "Added Value of Radiomics on Mammography for Breast Cancer Diagnosis: A Feasibility Study". *Journal of the American College of Radiology* 16 (Apr. 2019), 485–491.

[10] V. S. Parekh and M. A. Jacobs. "Integrated Radiomic Framework for Breast Cancer and Tumor Biology Using Advanced Machine Learning and Multiparametric MRI". *npj Breast Cancer* 3 (Nov. 2017), 43.

[11] B. S. Abunasser, M. R. J. AL-Hiealy, I. S. Zaqout, and S. S. Abu-Naser. "Breast Cancer Detection and Classification Using Deep Learning Xception Algorithm". *International Journal of Advanced Computer Science and Applications* 13 (2022).

[12] B. Abunasser, M. R. AL-Hiealy, I. Zaqout, and S. Abu-Naser. "Convolution Neural Network for Breast Cancer Detection and Classification Using Deep Learning". *Asian Pacific Journal of Cancer Prevention* 24 (Feb. 2023), 531–544.

[13] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin. "A Curated Mammography Data Set for Use in Computer-Aided Detection and Diagnosis Research". *Scientific Data* 4 (Dec. 2017), 170177.

[14] A. Dosovitskiy et al. *An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale*. June 2021.

[15] F. Manigrasso, R. Milazzo, A. S. Russo, F. Lamberti, F. Strand, A. Pagnani, and L. Morra. "Mammography Classification with Multi-View Deep Learning Techniques: Investigating Graph and Transformer-Based Architectures". *Medical Image Analysis* 99 (Jan. 2025), 103320.

[16] B. Q. Huynh, H. Li, and M. L. Giger. "Digital Mammographic Tumor Classification Using Transfer Learning from Deep Convolutional Neural Networks". *Journal of Medical Imaging* 3 (Aug. 2016), 034501.

[17] H. Li and X. Jiao. "Research on Machine Learning Classification Model of Mammography Images Based on Radiomics". *Journal of Taiyuan University of Technology* 53 (2022), 728–735.

[18] A. Padilla, O. Arponen, I. Rinta-Kiikka, and S. Pertuz. "Image Retrieval-Based Parenchymal Analysis for Breast Cancer Risk Assessment". *Medical Physics* 49 (2022), 1055–1064.

[19] H. M. Whitney, N. S. Taylor, K. Drukker, A. V. Edwards, J. Papaioannou, D. Schacht, and M. L. Giger. "Additive Benefit of Radiomics Over Size Alone in the Distinction Between Benign Lesions and Luminal A Cancers on a Large Clinical Breast MRI Dataset". *Academic Radiology* 26 (Feb. 2019), 202–209.

[20] G. Lu, R. Tian, W. Yang, R. Liu, D. Liu, Z. Xiang, and G. Zhang. "Deep Learning Radiomics Based on Multimodal Imaging for Distinguishing Benign and Malignant Breast Tumours". *Frontiers in Medicine* 11 (July 2024), 1402967.

[21] L. G. Falconi, M. Perez, W. G. Aguila, and A. Conci. "Transfer Learning and Fine Tuning in Breast Mammogram Abnormalities Classification on CBIS-DDSM Database". *Advances in Science, Technology and Engineering Systems Journal* 5 (2020), 154–165.

[22] M. A. Jones, N. Sadeghipour, X. Chen, W. Islam, and B. Zheng. "A Multi-stage Fusion Framework to Classify Breast Lesions Using Deep Learning and Radiomics Features Computed from Four-view Mammograms". *Medical Physics* 50 (Dec. 2023), 7670–7683.

[23] T. Mahmood, T. Saba, A. Rehman, and F. S. Alamri. "Harnessing the Power of Radiomics and Deep Learning for Improved Breast Cancer Diagnosis with Multiparametric Breast Mammography". *Expert Systems with Applications* 249 (Sept. 2024), 123747.

[24] V. M. Tiryaki and N. Tutkun. "Breast Cancer Mass Classification Using Machine Learning, Binary-Coded Genetic Algorithms and an Ensemble of Deep Transfer Learning". *The Computer Journal* 67 (Apr. 2024), 1111–1125.

[25] A. Carriero, L. Groenhoff, E. Vologina, P. Basile, and M. Albera. "Deep Learning in Breast Cancer Imaging: State of the Art and Recent Advancements in Early 2024". *Diagnostics* 14 (Apr. 2024), 848.

[26] X. Zhang, C. Liang, D. Zeng, X. Jiang, R. Zhong, Y. Lan, J. Ma, and L. Bai. "Pattern Classification for Breast Lesion on FFDM by Integration of Radiomics and Deep Features". *Computerized Medical Imaging and Graphics* 90 (June 2021), 101922.

[27] A. H. Yurttakal, H. Erbay, T. İkizceli, S. Karaçavuş, and C. Biçer. "Diagnosing Breast Cancer Tumors Using Stacked Ensemble Model". *Journal of Intelligent & Fuzzy Systems* 42 (Dec. 2021), 77–85.

[28] Y. Cui, Y. Li, D. Xing, T. Bai, J. Dong, and J. Zhu. "Improving the Prediction of Benign or Malignant Breast Masses Using a Combination of Image Biomarkers and Clinical Parameters". *Frontiers in Oncology* 11 (Mar. 2021), 629321.

[29] P. Lambin et al. "Radiomics: The Bridge between Medical Imaging and Personalized Medicine". *Nature Reviews Clinical Oncology* 14 (Dec. 2017), 749–762.

[30] C. N. Ogbonnaya, B. S. O. Alsaedi, A. J. Alhussaini, R. Hislop, N. Pratt, and G. Nabi. "Radiogenomics Reveals Correlation between Quantitative Texture Radiomic Features of Biparametric MRI and Hypoxia-Related Gene Expression in Men with Localised Prostate Cancer". *Journal of Clinical Medicine* 12 (Jan. 2023), 2605.

[31] V. Parekh and M. A. Jacobs. "Radiomics: A New Application from Established Techniques". *Expert Review of Precision Medicine and Drug Development* 1 (Mar. 2016), 207–226.

[32] A. Demircioğlu. "The Effect of Feature Normalization Methods in Radiomics". *Insights into Imaging* 15 (Jan. 2024), 2.

[33] Y. Liu, L. Xu, J. Wang, and C. Li. "A comprehensive review of publicly available mammography databases for computer-aided diagnosis research". *Frontiers in Oncology* 13 (2023), 1176474.

[34] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin. "A curated mammography data set for use in computer-aided detection and diagnosis research". *Scientific Data* 4 (2017), 170177.

[35] W. Lotter et al. "Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach". *Nature Medicine* 28 (2022), 1140–1150.

[36] H. Avcı and J. Karakaya. "A Novel Medical Image Enhancement Algorithm for Breast Cancer Detection on Mammography Images Using Machine Learning". *Diagnostics* 13 (Jan. 2023), 348.

[37] H. Li, M. Giger, O. Olopade, A. Margolis, L. Lan, and M. Chinander. "Computerized Texture Analysis of Mammographic Parenchymal Patterns of Digitized Mammograms 1". *Academic Radiology - ACAD RADIOL* 12 (July 2005), 863–873.

[38] A. Zwanenburg, S. Leger, M. Vallières, and S. Löck. "Image Biomarker Standardisation Initiative". *Radiology* 295 (May 2020), 328–338.

[39] A. Conti, A. Duggento, I. Indovina, M. Guerrisi, and N. Toschi. "Radiomics in Breast Cancer Classification and Prediction". *Seminars in Cancer Biology* 72 (July 2021), 238–250.

[40] W. Zhang, Y. Guo, and Q. Jin. "Radiomics and Its Feature Selection: A Review". *Symmetry* 15 (Oct. 2023), 1834.

[41] C. Cortes and V. Vapnik. "Support-Vector Networks". *Machine Learning* 20 (Sept. 1995), 273–297.

[42] S. Rizzo, F. Botta, S. Raimondi, D. Origgi, C. Fanciullo, A. G. Morganti, and M. Bellomi. "Radiomics: The Facts and the Challenges of Image Analysis". *European Radiology Experimental* 2 (Nov. 2018), 36.

[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger. Vol. 25. Curran Associates, Inc., 2012.

[44] R. M. Haralick, K. Shanmugam, and I. H. Dinstein. "Textural features for image classification". *IEEE Transactions on systems, man, and cybernetics* (2007), 610–621.

[45] M. M. Galloway. "Texture Analysis Using Gray Level Run Lengths". *Computer Graphics and Image Processing* 4 (June 1975), 172–179.

[46] G. Thibault, B. Fertil, C. Navarro, S. Pereira, P. Cau, N. Levy, J. Sequeira, and J.-L. Mari. "Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification" ().

[47] W. S. McCulloch and W. Pitts. "A Logical Calculus of the Ideas Immanent in Nervous Activity". *The Bulletin of Mathematical Biophysics* 5 (Dec. 1943), 115–133.

[48] F. Rosenblatt. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Review* 65 (1958), 386–408.

[49] M. Minsky and S. Papert. "An introduction to computational geometry". *Cambridge tiass., HIT* 479 (1969), 104.

[50] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning Representations by Back-Propagating Errors". *Nature* 323 (Oct. 1986), 533–536.

[51] B. Widrow and M. E. Hoff. "(1960) Bernard Widrow and Marcian E. Hoff, "Adaptive Switching Circuits," 1960 IRE WESCON Convention Record, New York: IRE, Pp. 96-104". *Neurocomputing, Volume 1*. Ed. by J. A. Anderson and E. Rosenfeld. The MIT Press, Apr. 1988, 126–134.

[52] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[53] C. M. Bishop. *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[54] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber. *Gradient flow in recurrent nets: the difficulty of learning long-term dependencies*. Tech. rep. A field guide to dynamical recurrent networks. Institut Dalle Molle D'intelligence Artificielle Perceptive, 1998.

[55] X. Glorot, A. Bordes, and Y. Bengio. "Deep sparse rectifier neural networks". *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2011, 315–323.

[56] L. Lu, Y. Shin, Y. Su, and G. E. Karniadakis. "Dying ReLU and Initialization: Theory and Numerical Examples". *Communications in Computational Physics* 28 (2020), 1671–1706.

[57] X. Qi, T. Wang, and J. Liu. "Comparison of Support Vector Machine and Softmax Classifiers in Computer Vision". *2017 Second International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*. Harbin: IEEE, Dec. 2017, 151–155.

[58] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi. "Convolutional Neural Networks: An Overview and Application in Radiology". *Insights into Imaging* 9 (Aug. 2018), 611–629.

[59] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-Based Learning Applied to Document Recognition". *Proceedings of the IEEE* 86 (Nov. 1998), 2278–2324.

[60] H. Mukherjee, S. Ghosh, A. Dhar, O. Sk, K. Santosh, and K. Roy. "Shallow Convolutional Neural Network for COVID-19 Outbreak Screening Using Chest X-rays". *Cognitive Computation* 16 (Feb. 2021).

[61] S. J. Pan and Q. Yang. "A Survey on Transfer Learning". *IEEE Transactions on Knowledge and Data Engineering* 22 (May 2010), 1345–1359.

[62] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. "How transferable are features in deep neural networks?" *Advances in neural information processing systems*. 2014, 3320–3328.

[63] Z. Li and D. Hoiem. *Learning without Forgetting*. Feb. 2017.

[64] Z. Wang, H. Ren, R. Lu, and W. Liu. "Soft-voting-based SVM-KNN algorithm and its application in big data modeling [C]". *7th International Workshop on Advanced Computational Intelligence and Intelligent Informatics*. 2021.

[65] GeeksforGeeks. *Stacking in Machine Learning*. Accedido: 29 de agosto de 2025. July 2025.

[66] I. A. Urbaniak. "Using Compressed JPEG and JPEG2000 Medical Images in Deep Learning: A Review". *Applied Sciences* 14 (Nov. 2024), 10524.

[67] Y. Gao, J. Lin, Y. Zhou, and R. Lin. "The Application of Traditional Machine Learning and Deep Learning Techniques in Mammography: A Review". *Frontiers in Oncology* 13 (Aug. 2023).

[68] A. Stefano et al. *Comparative Evaluation of Machine Learning-Based Radiomics and Deep Learning for Breast Lesion Classification in Mammography*. en. Mar. 2025.

[69] J. J. Van Griethuysen et al. "Computational Radiomics System to Decode the Radiographic Phenotype". *Cancer Research* 77 (Nov. 2017), e104–e107.

[70] K. He, X. Zhang, S. Ren, and J. Sun. *Deep Residual Learning for Image Recognition*. 2015.

[71] K. Simonyan and A. Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. Version Number: 6. 2014.

[72] N. Lauciello, E. Giovagnoli, G. Pasini, F. Bini, F. Marinozzi, G. Russo, and A. Stefano. "Mammography Classification: How Useful is Machine Learning? A Radiomics Study and Future Perspectives". *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*. Madrid, Spain: IEEE, June 2025, 119–120.

[73] S. K. Bhoi, P. Khandelwal, A. Tiwari, and A. Chugh. "Detection and classification of cancer in colon and skin samples using Haralick features and texture analysis". *Medical & Biological Engineering & Computing* 57 (2019), 2473–2486.

[74] A. S. Tagliafico, M. Piana, D. Schenone, R. Lai, A. M. Massone, and N. Houssami. "Breast cancer: Radiomics and machine learning for diagnostic and prognostic imaging". *European Journal of Radiology* 130 (2020), 109135.

[75] J. Wang, F. Kato, N. Oyama-Manabe, R. Li, Y. Cui, K. Tha, H. Yamashita, H. Shirato, and N. Kodama. "Breast lesion classification with multiparametric breast MRI using radiomics and machine learning". *European Radiology* 27 (2017), 4357–4368.