

Prototipo de un sistema de diagnóstico asistido por ordenador orientado a la localización de clusters de microcalcificaciones en mamografías.

Trabajo de grado para optar al título de Magíster en Ingeniería Eléctrica

Estudiante: Kevin Alejandro Hernández Gómez

Director: Álvaro Ángel Orozco Gutiérrez

Co-director: David Augusto Cárdenas Peña



Universidad Tecnológica de Pereira
Facultad de Ingenierías Eléctrica, Electrónica, Física y Ciencias de la
Computación
Maestría en Ingeniería Eléctrica
Grupo de Investigación en Automática
Pereira - Risaralda
2022

Índice

1. Resumen	1
2. Abstract	2
3. Introducción	3
4. Planteamiento del problema	4
5. Justificación	8
6. Objetivos	9
6.1. General	9
6.2. Específicos	9
7. Revisión del estado del arte	10
7.1. Eliminación de artefactos y remoción del musculo pectoral	10
7.2. Detección de clusters de microcalcificaciones	11
7.3. Clasificación de clusters de microcalcificaciones	12
8. Marco Teórico	14
8.1. Eliminación de ruido y supresión de artefactos	15
8.1.1. Realce de contraste por corrección gamma adaptativa con ponderación distribuida	15
8.1.2. Segmentación del tejido mamario	16
8.2. Remoción del musculo pectoral	18
8.2.1. Segmentación del musculo pectoral con K-medias	19
8.2.2. Corrección del contorno con aproximación polinomial	20
8.3. Detección y localización de microcalcificaciones	22
8.3.1. Detección de MC con CNN	22
8.3.2. Realce de contraste, segmentación y filtrado de MC	24
8.4. Clasificación de microcalcificaciones según su categoría BI-RADS	26
8.4.1. Escala BI-RADS	26
8.4.2. Extracción de características	27
8.4.3. Redes Neuronales Artificiales	27
8.4.4. Sistema de recuperación de imágenes de microcalcificaciones	28

9. Marco experimental	30
9.1. Bases de datos	30
9.2. Resultados de las pruebas de eliminación de artefactos	31
9.3. Resultados de las pruebas de remoción del musculo pectoral	32
9.4. Resultados de las pruebas de detección y localización de MC	34
9.5. Resultados de las pruebas de clasificación de MC según su categoría BI-RADS	36
9.6. Diseño de la interfaz del sistema DAO	37
10. Conclusiones	41
11. Resultados académicos	44
12. Agradecimientos	45
13. Bibliografía	46

Índice de figuras

1.	Artefactos presentes en una imagen mamografica tipica.	14
2.	Localización del umbral de Otsu a partir del histograma de la imagen J	17
3.	Pasos en la eliminación de artefactos	18
4.	Pasos uno y dos de la remoción del musculo pectoral.	19
5.	Pasos tres y cuatro para la remoción del musculo pectoral.	21
6.	Aproximación polinomial del contorno del musculo pectoral.	21
7.	Arquitectura básica de una CNN.	23
8.	Pasos para localización de MC.	24
9.	Resultado final del procesamiento, donde se muestra una mamografía sin artefactos, sin musculo pectoral y con las MC localizadas.	25
10.	Arquitectura básica de una RNA.	28
11.	Esquema metodológico del CBIR.	29
12.	Ejemplos segmentación correcta e incorrecta.	31
13.	Ejemplos de remoción del musculo pectoral.	33
14.	Matriz de confusión resultante de la etapa de entrenamiento.	35
15.	Matriz de confusión modelo de ensamble de RNA.	36
16.	Panel inicial del sistema DAO.	38
17.	Panel de información del paciente.	39
18.	Panel de recuperación de información.	40

Índice de cuadros

1.	Especificaciones de las BD.	30
2.	Resultados de las pruebas de eliminación de artefactos comparados con otros métodos propuestos en la literatura.	32
3.	Resultados del método de remoción del musculo pectoral comparados con otros resultados reportados en la literatura.	33
4.	Sensibilidad y tasa de FP por imagen comparados con otros resultados reportados en la literatura.	35
5.	Resultados de la clasificación de MC según la escala BI-RADS comparados con otros resultados reportados en la literatura.	37

1. Resumen

En este trabajo se presenta la construcción metodológica para la implementación de un prototipo de aplicativo software que sirva como herramienta de apoyo al diagnóstico de cáncer de mama, a partir de las diferentes técnicas de procesamiento de imágenes y modelos de aprendizaje supervisado y no-supervisado. Tiene como aporte fundamental el hecho de que es una metodología que acopla diferentes etapas de procesamiento bastante robustas que permiten hacer un tratamiento desde la imagen mamográfica en crudo hasta la recomendación final dada por el sistema (End-to-End). En particular se consideró la técnica de realce de contraste de corrección gamma adaptativa con ponderación distribuida (AGCWD) y binarización de Otsu para la segmentación del tejido mamario, el segmentador K-means para la identificación del musculo pectoral, una red neuronal convolucional (CNN) para la localización de microcalcificaciones, un ensamble de redes neuronales artificiales (RNA) responsable de la clasificación y del proceso de búsqueda de imágenes similares. Además, se usó la librería tkinter para la implementación de la interfaz gráfica de usuario (GUI) en Python. Para la validación de la metodología se usaron dos bases de datos, The Mammographic Image Analysis (mini-MIAS) y The Digital Database for Screening Mammography (DDSM). Los resultados obtenidos reflejan que esta metodología mejora sustancialmente el rendimiento en la eliminación de artefactos (99.78 %), la precisión en la remoción del musculo pectoral (92.14 %), la reducción de falsos positivos en la detección de microcalcificaciones (0.47 por imagen), y aumento en el acierto en la clasificación según el estándar BI-RADS (82 %) en comparación a otros trabajos en el estado del arte.

2. Abstract

This work presents the methodological construction for the implementation of a prototype software application that serves as a support tool for breast cancer diagnosis, based on different image processing techniques and supervised and unsupervised learning models. Its fundamental contribution is the fact that it is a methodology that couples different processing stages quite robust that allow a treatment from the raw mammographic image to the final recommendation given by the system (End-to-End). In particular, the contrast enhancement with adaptive gamma correction weighting distribution (AGCWD) and Otsu binarization technique was considered for the segmentation of breast tissue, the K-means segmenter for the identification of pectoral muscle, a convolutional neural network (CNN) for the localization of microcalcifications, an assembly of artificial neural networks (ANN) responsible for the classification and the search process of similar images. In addition, the tkinter library was used for the implementation of the graphical user interface (GUI) in Python. Two databases, The Mammographic Image Analysis (mini-MIAS) and The Digital Database for Screening Mammography (DDSM), were used to validate the methodology. The results obtained reflect that this methodology substantially improves the performance in the elimination of artifacts (99.78%), the accuracy in the removal of the pectoral muscle (92.14%), the reduction of false positives in the detection of microcalcifications (0.47 per image), and the increase in the accuracy in the classification according to the BI-RADS standard (82%) in comparison to other works in the state of the art.

3. Introducción

El cáncer de mama es una de las principales causas de morbilidad y mortalidad en el mundo, en la actualidad son 2,3 millones de casos cada año y 685,000 muertes en el 2020 [1]. Teniendo presente que esta tendencia sigue en aumento, se hace imprescindible la implementación de estrategias de detección temprana de cáncer de mama, de modo que los tratamientos puedan iniciarse en la mayor brevedad y se contribuya así a la reducción de la tasa de mortalidad debida a esta enfermedad. Para diagnosticar el cáncer de mama en una etapa temprana, se ha recurrido a la detección visual de clusters de microcalcificaciones (MC) a partir de imágenes mamográficas, los cuales a menudo son de origen benigno y un indicador común para el diagnóstico de esta enfermedad [2], además de que se ha estandarizado una terminología (categorías BI-RADS) que permite un mejor análisis y evaluación de los hallazgos en mamografías [12]. Sin embargo, la detección y clasificación de clusters de MC a partir de mamografías es particularmente difícil debido a la gran variedad de composición, anatomía altamente texturizada y estructuras similares a las MC presentes en el tejido mamario [3].

Por lo tanto, es evidente que existe la necesidad de reducir los problemas inherentes al diagnóstico de esta enfermedad a partir de imágenes mamográficas, por este motivo se ha optado por el uso de técnicas de procesamiento de imágenes para acondicionar la imagen mejorando la calidad de la misma [4], así mismo se han planteado metodologías para detectar clusters de MC, empleando técnicas de aprendizaje de máquina [5]. A pesar de esto, sigue existiendo una marcada desconfianza por parte de la comunidad médica en cuanto al uso de estas herramientas para apoyar la etapa del diagnóstico. En consecuencia, con este proyecto se busca implementar un prototipo de sistema de diagnóstico asistido por ordenador (DAO) que permita identificar, localizar y clasificar clusters de MC en mamografías, contribuyendo de esta manera a la detección temprana del cáncer de mama y a la integración de técnicas de procesamiento de imágenes y aprendizaje de máquina en el proceso de diagnóstico médico.

4. Planteamiento del problema

El cáncer de mama es una enfermedad progresiva en la que las células comienzan a dividirse de forma descontrolada en el tejido mamario, resultando en un bulto o masa. Hay muchos tipos diferentes de cáncer de mama, el más común es el carcinoma ductal y comienza en el revestimiento de los conductos que llevan la leche materna al pezón, una vez las células mueren se acumulan en pequeñas manchas de calcio o microcalcificaciones [6]. El cáncer de mama puede propagarse cuando las células cancerosas entran en los vasos linfáticos y comienzan a dividirse, por lo que existe una mayor probabilidad de que las células viajen a través del sistema linfático y se diseminen (metástasis) a otros órganos [7]. En promedio el cáncer de mama afecta a 2,3 millones de mujeres cada año, siendo la enfermedad mas frecuente entre mujeres, también es su principal causa de muerte, se estima que, en el año 2020, murieron 685,000 mujeres, es decir, aproximadamente el 16 % de todas las muertes por cáncer. Además, las cifras no son muy alentadoras ya que la tasa de mujeres diagnosticadas con cáncer de mama sigue en aumento [1]. Por este motivo la detección temprana del cáncer de mama es primordial, puesto que los pacientes pueden tener acceso anticipado a servicios de diagnóstico efectivos, lo que permite un tratamiento más eficaz para ser utilizado y reducir los riesgos de muerte por cáncer de mama [2]. En la actualidad existen diferentes técnicas para realizar exámenes médicos que permitan la detección temprana del cáncer de mama, modalidades como: tomografía de impedancia eléctrica, tomografía de difusión óptica, radiometria de microondas, ultrasonido e imágenes de resonancia biomagnética y magnética son algunas de las técnicas comúnmente usadas [8], sin embargo estas modalidades todavía enfrentan grandes desafíos y limitaciones que necesitan ser superados antes de que puedan ser introducidos como herramientas de uso clínico a gran escala, como por ejemplo el nivel de acierto, ya que uno de cada cuatro diagnósticos es incorrecto, además del alto costo, la seguridad y conformidad del paciente durante el examen también se ven comprometidas [9].

Por otro lado, tanto la Organización Mundial de la Salud como la Asociación Americana de Cáncer consideran a la mamografía como el procedimiento estándar de diagnóstico por imágenes con más alta sensibilidad para detectar signos tempranos de cáncer de mama, se ha demostrado que la detección de cáncer de mama mediante mamografía puede reducir la tasa de mortalidad hasta un 20 % [6, 10]. Existen varios indicadores de cáncer de mama presentes en las mamografías como masas, distorsión arquitectónica, MC y la asimetría bilateral. Los clusters de MC pueden ser el único indicador temprano del cáncer de mama y son casos de carcinoma ductal en casi el 93 %. Por lo tanto, la detección de clústeres de MC es vital en el pronóstico del cáncer

de mama [11]. Para estandarizar el análisis, la evaluación y el manejo de hallazgos mamográficos en este tipo de imágenes, el Colegio Americano de Radiología desarrolló una terminología estandarizada denominada Breast Imaging Reporting and Data System (BI-RADS). Esta escala estándar incluye además categorías para los hallazgos, los cuales son clasificados de acuerdo con su probabilidad de malignidad [12].

Sin embargo, los radiólogos a menudo están expuestos a una alta carga de mamografías a analizar y a extensas jornadas laborales, lo que provoca una elevada tasa de falsos negativos (aproximadamente el 20 %) y el diagnóstico erróneo debido a la fatiga, dado que la inspección visual es una tarea extremadamente exigente y que consume mucho tiempo, a esto se suman las dificultades inherentes asociadas con la lectura de imágenes de mamografía, es decir, la complejidad de la estructura del tejido mamario y la naturaleza sutil del cáncer durante las primeras etapas de desarrollo [3]. La identificación de MC es una tarea difícil, no sólo por la gran variedad de composición y anatomía mamaria altamente texturizada, también por el bajo contraste inherente de las mamografías y, en algunos casos, el tamaño impalpable de las MC (0,5 – 2 mm de diámetro), también diferentes tipos de estructuras como las estructuras curvilíneas pueden aparecer junto con las MC lo que dificultan su identificación y clasificación en en la escala BI-RADS [5].

Por lo tanto, es necesaria una herramienta de diagnóstico asistido por ordenador que pueda detectar y clasificar clusters de MC o mamografías normales con precisión, minimizar la carga, mejorar el rendimiento del radiólogo y proporcionar una segunda opinión durante el examen. El uso de DAO ha sido propuesto por muchos investigadores y es adoptado por muchos centros de imágenes mamarias, y los estudios han demostrado que el rendimiento del radiólogo puede mejorarse mediante la utilización de DAO en su evaluación [4]. No obstante, para detectar anomalías utilizando sistemas DAO, necesitamos eliminar algunas regiones no deseadas de la mamografía, como ruido, artefactos radiopacos y el musculo pectoral, así como realzar el contraste de las MC sobre otros tipos de estructuras presentes en la mama, esto con el objetivo de maximizar la eficiencia del sistema, ya que todos estos elementos tienen características texturales que puede conducir fácilmente a una tasa de falsos positivos (FP) y un diagnóstico erróneo de cáncer de mama [13].

Por lo general los sistemas DAO están conformados por dos etapas, una encargada del preprocesamiento de la mamografía, cuyo objetivo es lograr una correcta segmentación de la mama, y otra dedicada a la detección cuyo objetivo es identificar de forma precisa donde están ubicados los clusters de MC. Sin embargo, la implementación

de ambas etapas continúa siendo un reto ya que la determinación del contorno de la mama se ve afectada tanto por el ruido, como por las etiquetas inducidas por el escáner y los artefactos provocados por la posición de la mama durante el proceso de digitalización de la imagen, además de las formas complejas del contorno del musculo pectoral, la alta densidad de algunos tejidos en la mama, la apariencia de pliegues auxiliares, y la superposición del musculo pectoral con el tejido fibroglandular hacen del preprocesamiento de la mamografía una etapa bastante compleja [4, 14]. Por otro lado, el problema desafiante al que se enfrenta la etapa de detección de MC es la ocurrencia frecuente de FP. Esto se debe a que la respuesta del detector es susceptible a los patrones de imagen locales como las fuertes variaciones del fondo que pueden introducir estructuras similares a las MC, y niveles de ruido que a menudo apenas son superados por las señales. Además, las estructuras lineales de la mamografía como conductos, vasos sanguíneos, tejidos fibrosos, ligamentos y puntos de intersección pueden producir una apariencia de textura característica haciendo que estos sean más similares a las MC que al ruido, lo que puede causar una detección falsa [15].

De la misma forma los sistemas DAO enfocados a la clasificación de MC en categorías BI-RADS tienden a fallar a la hora de discriminar la categoría BI-RADS 3 de las demás [16, 17] lo que representa un problema desafortunado ya que la presencia y estabilidad a lo largo del tiempo de esta categoría, hace que el hallazgo sea catalogado como benigno o, por el contrario, si se detecta un cambio sustancial en la morfología y densidad, los especialistas recomiendan hacer una biopsia de los tejidos [18]. Repetir los exámenes diagnósticos es menos costoso si se compara con una biopsia; sin embargo, examinar repetidamente a un paciente puede generar en éste ansiedad, debido a la posibilidad de que los hallazgos no sean concluyentes y que permanezca, de forma latente, un posible diagnóstico de cáncer que no esté siendo tratado [19]. Esta situación indeseable puede llegar a dilatarse en algunos casos hasta por más de dos años, tiempo en el cual la conformidad del paciente disminuye y un diagnóstico no concluyente se vuelve intolerable [17]. Por esto, los diagnósticos basados en mamografías categorizadas como BI-RADS 3 son los que más problemas pueden llegar a presentar.

Múltiples metodologías se han planteado [20, 21, 22, 23, 24, 25] con el objetivo de solucionar los inconvenientes previamente mencionados, no obstante las técnicas de segmentación propuestas no garantizan una segmentación suave del contorno de la mama, además de no contar con un método que ajuste la curvatura del contorno del musculo pectoral para conseguir una segmentación mucho más precisa. Por otro lado los trabajos propuestos en [26, 5, 27] para la detección de MC tienen el principal in-

conveniente de realizar la búsqueda en regiones de interés previamente seleccionadas de forma manual y en algunos casos la alta sensibilidad de los algoritmos empleados conlleva a una alta tasa de FP imposibilitando diagnósticos mucho mas precisos. Además en cuanto a la clasificación de MC se han desarrollado trabajos como [34, 35, 31], sin embargo, ninguno de estos enfoques de clasificación considera a las categorías BI-RADS por lo que solo discriminan MC benignas de malignas, cuando es evidentemente necesario realizar una clasificación de tipo multiclase que considere dicho estándar, por lo que las problemáticas metodológicas que giran entorno al preprocesamiento de imágenes mamográficas como a la localización y clasificación de clusters de MC continúan sin resolverse.

Basados en los problemas antes mencionados se plantea la siguiente pregunta de investigación: ¿cómo desarrollar una metodología que permita implementar un sistema de diagnóstico asistido por ordenador basado en las técnicas de procesamiento de imágenes y aprendizaje de maquina para el preprocesamiento, la identificación de clusters de MC y su clasificación en la escala BI-RADS?.

5. Justificación

El cáncer de mama es la causa mas frecuente de muerte por cáncer entre mujeres con 2,3 millones de casos cada año y 685,000 muertes en el 2020 [1]. Por esto, diagnosticar y tratar el cáncer de mama en una etapa temprana es fundamental, dado que un diagnostico oportuno permite un tratamiento más eficaz para ser utilizado y reducir sustancialmente la tasa de mortalidad por esta enfermedad. Existen características observadas clínicamente, entre ellas las MC, las cuales a menudo son de origen benigno y un indicador común para el diagnóstico de cáncer de mama, por lo que son de especial importancia en el diagnostico temprano de esta enfermedad [2].

Sin embargo, la detección de clusters de MC a partir de mamografías está sujeto en gran parte a la experticia e interpretación del radiólogo, en muchos casos esto conlleva al sobrediagnóstico del paciente y el aumento de la tasa de FP [3]. Por ello, es necesaria una herramienta de diagnóstico asistido por ordenador que pueda detectar clusters de MC o mamografías normales con precisión, minimizar la carga del especialista reduciendo el esfuerzo requerido para realizar el diagnóstico, mejorar el rendimiento del radiólogo dotándolo de una mayor certeza en su toma de decisiones, y proporcionar una segunda opinión durante el examen. Además, se ha demostrado que los sistemas DAO, en su uso como apoyo al diagnóstico de cáncer de mama, mejoran la sensibilidad de los radiólogos del 73,5 % al 87,4 % [28], esclareciendo así el diagnóstico de una enfermedad evitable.

Por lo anterior, con este proyecto se busca desarrollar una metodología que permita identificar, localizar y clasificar clusters de MC en mamografías, contribuyendo de esta manera a la detección temprana del cáncer de mama y a la integración de técnicas de procesamiento de imágenes y aprendizaje de maquina en el proceso de diagnóstico médico. Gracias a la implementación de un sistema DAO el especialista podría realizar un diagnóstico más preciso reduciendo las tasas de FP y los diagnósticos no concluyentes en etapas prequirúrgicas; de esta forma se reducirían las biopsias innecesarias a la vez que se mejoraría la percepción del paciente sobre el diagnóstico y del propio personal sanitario sobre la utilidad de las herramientas de última generación.

6. Objetivos

6.1. General

Desarrollar un prototipo de sistema de diagnóstico asistido por ordenador basado en técnicas de procesamiento de imágenes y aprendizaje de máquina para la localización de clusters de microcalcificaciones en mamografías y su clasificación en la escala BI-RADS.

6.2. Específicos

- Desarrollar una metodología de segmentación de mama y remoción del musculo pectoral empleando técnicas de procesamiento de imágenes para disminuir el ruido, eliminar artefactos y mejorar el contraste.
- Desarrollar una metodología para la detección y localización de clusters de microcalcificaciones en mamografías, empleando técnicas de aprendizaje de máquina robustas a las diferentes variaciones del tejido mamario.
- Desarrollar una metodología para la recuperación de imágenes de microcalcificaciones similares basado en un enfoque de clasificación según su categoría BI-RADS.

7. Revisión del estado del arte

Los sistemas DAO para la identificación y clasificación de clusters de MC pueden reducir la elaboración de exámenes diagnósticos o biopsias innecesarias, que no solo incrementan los costos del diagnóstico, sino que también resultan incómodos y estresantes para los pacientes. Además, dicha herramienta disminuiría la incertidumbre en el diagnóstico de cáncer de mama, reduciendo el número de FP que puedan presentarse en este, sirviendo, así como soporte al médico radiólogo [2]. No obstante, existen problemas inherentes a la identificación y clasificación de MC a partir de imágenes de mamografías, como, por ejemplo, el ruido, las etiquetas, los artefactos, las formas complejas del contorno del musculo pectoral, la alta densidad de algunos tejidos en la mama [4, 14], las fuertes variaciones del fondo, las estructuras como conductos, vasos sanguíneos, tejidos fibrosos, ligamentos y puntos de intersección, pueden afectar el rendimiento del sistema DAO incrementando la tasa de FP [15].

A continuación, se presentan los trabajos mas recientes encontrados en la literatura en relación a las etapas de preprocesamiento, detección y clasificación de clusters de MC.

7.1. Eliminación de artefactos y remoción del musculo pectoral

Una etapa fundamental en la implementación de sistemas DAO para la detección y clasificación de MC es la fase de preprocesamiento de la mamografía, esta consiste en la eliminación de artefactos y la remoción del musculo pectoral con la finalidad de segmentar la mama. Diversas metodologías se han propuesto [20, 21, 22, 23], estandarizando una serie de pasos que por lo general consisten en realizar una etapa de filtrado usando para ello un filtro de media o gaussiano, luego binarizar usando el método de umbralización de Otsú, posteriormente aplicar una serie de operaciones morfológicas para eliminar artefactos, y finalmente tomar el objeto que este compuesto por el mayor número de pixeles (ya que este representaría la región mamaria). Sin embargo, la mayoría de estos métodos no proveen medidas estadísticas de su rendimiento y omiten los pasos necesarios para garantizar una segmentación suave del contorno de la mama [4].

Otras metodologías [29, 30] realizan una etapa intermedia de realce de contraste, antes de binarizar la imagen usando para ello la técnica de corrección Gamma o la técnica de ecualización de histogramas adaptativa por contraste limitado (CLAHE),

el problema principal con estos métodos de realce de contraste, en general, es el hecho de que no sólo se incrementa el contraste de la mama (considerado como primer plano), también se aumenta el contraste del fondo. Esto puede conducir a un aumento de intensidad de los artefactos inducidos por el escáner, lo que resulta en posibles dificultades para los enfoques basados en umbrales. Además de que muchas técnicas de segmentación de la mama tienen una limitación cuando las etiquetas aparecen demasiado cerca del límite del contorno mamario, considerando esto como la principal dificultad para lograr una segmentación satisfactoria [4].

Por otro lado, [24, 25] propusieron usar el algoritmo de crecimiento de regiones (RG) para segmentar el musculo pectoral basándose en el hecho de que esta región está localizada cerca del pecho, en la parte superior izquierda de la imagen. De esta forma asignan la semilla del algoritmo, y posteriormente usaron una línea para aproximar el contorno del musculo pectoral. Sin embargo, la gran mayoría de las metodologías de segmentación del musculo pectoral basadas en el algoritmo RG tienen el principal inconveniente de la elección de la posición de las semillas, ya que dependen de la ubicación esperada del musculo pectoral para su segmentación, estos métodos pueden tener dificultades en los casos en el que la imagen no está alineada correctamente o en casos en los que el músculo pectoral no está contenido dentro de esta región esperada. Además, los métodos de segmentación que asumen que el contorno del musculo pectoral tiene forma triangular pueden fallar ya que no es conveniente realizar esta aproximación debido a que en la gran mayoría de los casos clínicos reales se presenta una curvatura del contorno del musculo pectoral diferente a la de una recta [4].

7.2. Detección de clusters de microcalcificaciones

Otra etapa fundamental de los sistemas DAO para la detección de clusters de MC, es la fase de procesamiento de la mamografía, cuyo propósito es determinar si una mamografía contiene o no MC, y si es el caso determinar la posición de las mismas. Trabajos como [15, 11, 3] plantean metodologías en las que se realiza sustracción del fondo atenuando las estructuras irrelevantes usando técnicas de realce de contraste y operaciones morfológicas, también se realiza una umbralización para segmentar los candidatos a MC, y finalmente las MC son agrupadas en cluster con algoritmos como K-means, K-NN, entre otros. Sin embargo, algunos detectores están diseñados solo para la identificación de regiones sospechosas (ROI), y no especifican las posiciones individuales de cada MC, y por lo general en muchos de estos trabajos todos los casos de estudio contenían al menos una MC, cuando podría ser deseable incluir en

la evaluación de la metodología casos normales (sin MC). Además, la sensibilidad de muchos métodos depende de la elección de parámetros, como los criterios de agrupación de MC, que puede afectar al número obtenido de clusters de MC [26].

Otras metodologías como las propuestas en [26, 5, 27] basados en algoritmos de aprendizaje supervisado, proponen detectar MC, usando las características de las misma, obtenidas por medio de redes neuronales convolucionales o descomposición wavelet, posteriormente se usan clasificadores como el perceptrón multicapa o máquina de soporte vectorial para discriminar entre las regiones que contienen MC y las que no. Sin embargo, la mayoría de los enfoques mencionados en la literatura no localizan de forma automática los clusters de MC, y clasifican manualmente las regiones seleccionadas como clusters de MC o regiones normales. Además, estas metodologías carecen de una localización automática de las regiones sospechosas, de hecho, hacen uso de regiones seleccionadas manualmente. El principal inconveniente al que se enfrentan estas metodologías es la detección de objetos lineales, ya que es necesario para evitar errores de detección vinculados a nodos de la red vascular del tejido mamario, ya que todos los objetos candidatos con formas, tamaños y apariencias similares a las MC pueden inducir una alta tasa de FP [15].

7.3. Clasificación de clusters de microcalcificaciones

Los sistemas DAO en el contexto de la recuperación de imágenes basados en contenido (CBIR) tienen como objetivo entregar un conjunto de imágenes similares a una imagen consultada por un especialista [31], para esto se suele realizar la recuperación de imágenes guiada por un sistema de clasificación, con el propósito de incrementar la tasa de acierto en la recuperación de imágenes, disminuir el tiempo de búsqueda del sistema y en algunos casos entregar información extra como soporte al diagnostico (la clase predicha) [32, 33]. Diferentes estudios se han desarrollado en [34, 35, 31], donde se proponen metodologías de clasificación usando maquinas de soporte vectorial y clasificadores lineales basados en regresión logística, utilizando características de tipo morfológico y textural como vectores de representación, con el objetivo de servir de apoyo a un sistema CBIR y mejorar la precisión de las imágenes recuperadas. Sin embargo, estos enfoques de clasificación binario solo consideran MC benignas y malignas, cuando es necesario y pertinente considerar un enfoque multiclase determinado por las categorías BI-RADS, ya que esto permitiría al especialista realizar un diagnostico mucho mas preciso [36].

La investigación más reciente en relación a la clasificación de MC según su categorías

BI-RADS fue [16], donde se utilizó un modelo de red neuronal artificial (RNA) con dos capas ocultas, una función de activación sigmoidea y 10 unidades en cada capas oculta. La RNA se entrenó con vectores de características que contenían 10 descriptores morfológicos. Sin embargo, los resultados obtenidos no fueron satisfactorios ya que en términos de acierto en la clasificación solo se obtuvo un rendimiento del 71 % siendo insuficiente para un sistema CAD que pretende integrarse como herramienta de apoyo al diagnostico medico en casos clínicos reales.

8. Marco Teórico

Las imágenes mamográficas deben ser acondicionadas antes de ser procesadas por cualquier sistema DAO para tareas de clasificación o detección. Por lo general la etapa de preprocesamiento incluye: reducción de ruido, eliminación de artefactos, realce de contraste, segmentación de la mama y remoción del musculo pectoral [37]. Como se muestra en la Figura 1, el musculo pectoral constituye una región predominante de gran densidad en la mayoría de imágenes mamográficas con proyección oblicua mediolateral (MLO), las cuales pueden afectar los resultados de una posterior etapa de procesamiento, no solo en términos de rendimiento si no también en velocidad de análisis debido al tamaño significativo del área a analizar [38].

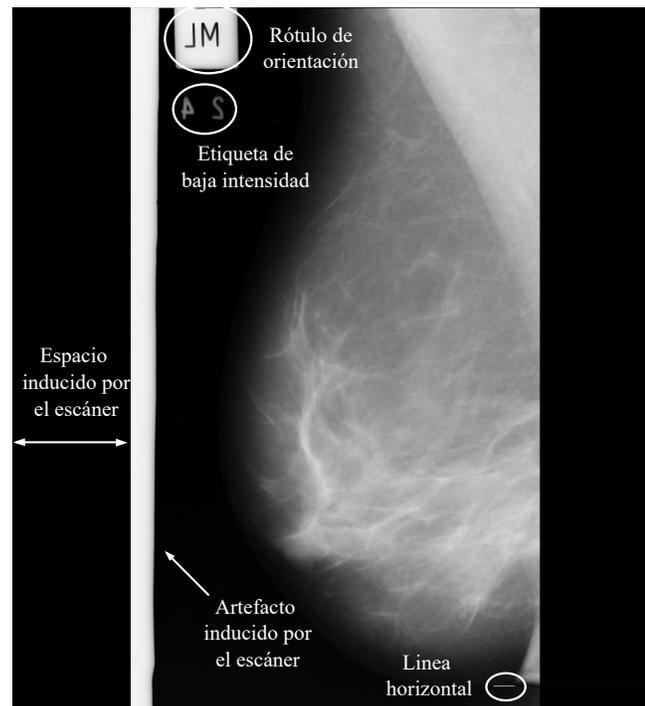


Figura 1: Artefactos presentes en una imagen mamografica tipica.

A continuación, se presenta el marco teórico de las técnicas empleadas para el acondicionamiento de la mamografía, así como los métodos usados para llevar a cabo la implementación de cada uno de los módulos del sistema DAO.

8.1. Eliminación de ruido y supresión de artefactos

Sea $\mathbf{I}_{RGB} \in \mathcal{N}^{h \times w \times 3}$ la imagen mamográfica original, de tres canales (RGB), donde h y w son la altura y el ancho de la imagen, se considera realizar una conversión a escala de grises debido a que esta escala es computacionalmente más eficiente y conveniente a la hora de usar modelos de aprendizaje automático [39]. La conversión a escala de grises se consigue promediando el valor del pixel (RGB) por medio de la expresión: $I_G(x, y) = \frac{1}{3} (I_{RGB}(x, y)_R + I_{RGB}(x, y)_G + I_{RGB}(x, y)_B)$. Por otro lado el ruido en las mamografías se entiende conceptualmente como fluctuaciones aleatorias que afectan el color o el brillo de las imágenes, se pueden presentar por diferentes factores, en muchos casos como parte del proceso de adquisición de las imágenes, por esto se propone un filtro de mediana con un tamaño de ventana (\mathbf{w}) de 3×3 para suavizar las imágenes, reduciendo el cambio brusco de intensidad, y eliminando las líneas verticales y horizontales inducidas por el escáner [40], la ecuación que permite realizar el filtrado es la siguiente: $I(x, y) = \text{Mediana} \{I_G(x - i, y - j), (i, j) \in \mathbf{w}\}$, donde (i, j) son los elementos de la ventana \mathbf{w} .

8.1.1. Realce de contraste por corrección gamma adaptativa con ponderación distribuida

En orden de segmentar el tejido mamario, es necesario encontrar el contorno de la mama, esto se convierte en una tarea complicada debido a que la intensidad de los pixeles cerca del contorno es similar a la de los pixeles del fondo, es por esto que es conveniente realzar el contraste de la imagen para incrementar la intensidad de los pixeles pertenecientes a la mama y así facilitar el proceso de segmentación de la misma. Se propone usar la técnica de corrección gamma adaptativa con ponderación distribuida (AGCWD) [38]. La corrección gamma puede representarse como:

$$J(x, y) = I_{max} \left(\frac{I(x, y)}{I_{max}} \right)^\gamma \quad (1)$$

Donde I_{max} es el valor de la intensidad máxima de la imagen, $J(x, y)$ son los valores nuevos de intensidad ($\mathbf{J} \in \mathcal{N}^{h \times w}$) y $\gamma = 1 - F_w(I)$ es un parámetro adaptativo, este método incrementa el valor de los pixeles de baja intensidad, evitando una degradación significativa en el valor de los pixeles de alta intensidad. La introducción de la distribución ponderada modifica el histograma y reduce los efectos negativos. La función de distribución ponderada se define como:

$$f_w(I) = f_{max} \left(\frac{f(I) - f_{min}}{f_{max} - f_{min}} \right)^\alpha \quad (2)$$

Donde α es un parámetro de ajuste, f_{max} y f_{min} son los valores máximos y mínimos de la función densidad de probabilidad. La función de distribución acumulada modificada se define como:

$$F_w(I) = \sum_{k=0}^I \frac{f_w(k)}{\sum f_w} \quad (3)$$

Donde la suma de la función de densidad modificada esta definida como: $\sum f_w = \sum_{k=0}^I f_w(k)$. La intensidad de los pixeles del fondo no es igual a cero, pero por lo general tienen un valor pequeño (entre 5 a 15 dependiendo de la imagen). En orden de evitar incrementar la intensidad de estos pixeles, se busca el primer valor diferente de cero del lado derecho e izquierdo de la imagen, el pixel con la menor intensidad es almacenado en la variable t , y si el valor es lo suficientemente pequeño ($t \leq 10$), la intensidad de todos los pixeles de la imagen es reducida un valor de t , posteriormente se aplica el método AGCWD como se muestra en la Figura 3a.

8.1.2. Segmentación del tejido mamario

Con el fin de segmentar el tejido mamario y así eliminar los artefactos mostrados en la Figura 1; una vez es aplicado el realce de contraste se opta por binarizar la imagen con el método de umbralización de Otsu, siendo el método más popular ya que produce buenos resultados en imágenes mamograficas con grandes variaciones de intensidad (imagen aproximadamente bimodal) [4], el umbral seleccionado es aquel que minimice la varianza intraclase, y se selecciona automáticamente del histograma de la imagen, en donde es posible evidenciar claramente dos picos que representan los diferentes rangos de valores de intensidad (Figura 2). El resultado de la binarización se muestra en la Figura 3b ($\mathbf{B} \in \mathcal{N}^{h \times w}$).

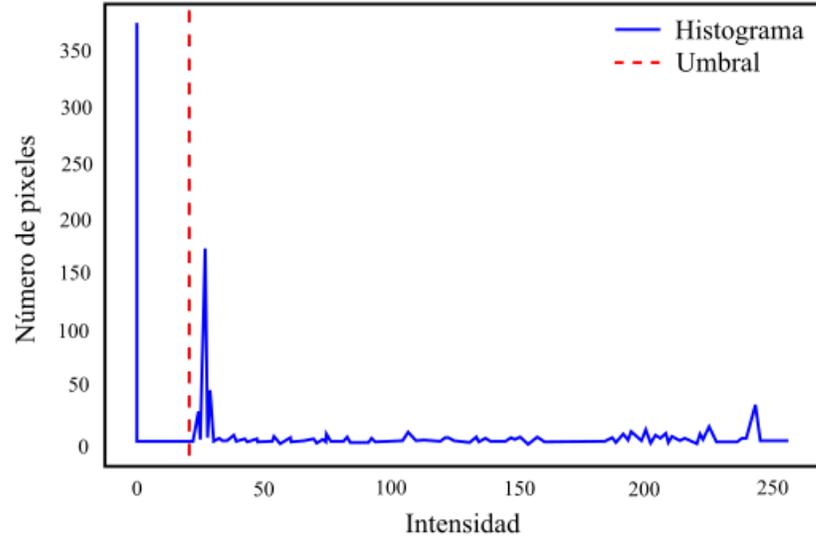


Figura 2: Localización del umbral de Otsu a partir del histograma de la imagen **J**.

Posteriormente, se remueven las franjas blancas inducidas por el escáner que en ocasión rodean la imagen mamográfica, esto se logra haciendo uso de una máscara binaria ($\mathbf{M} \in \mathcal{N}^{h \times w}$), la cual consiste en un marco con un ancho de $\lceil \frac{w}{15} \rceil$ píxeles, donde $\lceil \cdot \rceil$ es el operador parte entera [41]. Para eliminar las franjas se realiza el producto Hadamard (\otimes) entre \mathbf{B} y \mathbf{M} como se puede ver en la Figura 3c. Finalmente se empleó el método de componentes conectados el cual consiste en identificar las regiones blancas de la Figura 3c, la región más grande (área mayor) representa el área de interés ($\mathbf{R} \in \mathcal{N}^{h \times w}$) y el resto de las regiones son descartadas (Figura 3d); luego se realiza la operación $\mathbf{R} \otimes \mathbf{J}$ para extraer la región correspondiente al tejido mamario de la imagen original y después se reduce la imagen a un rectángulo que circunscribe la mama (BoundingBox de \mathbf{R}) dando como resultado la imagen preprocesada (Figura 3e) $\mathbf{P} \in \mathcal{N}^{h' \times w'}$, donde h' y w' son el alto y el ancho del BoundingBox.

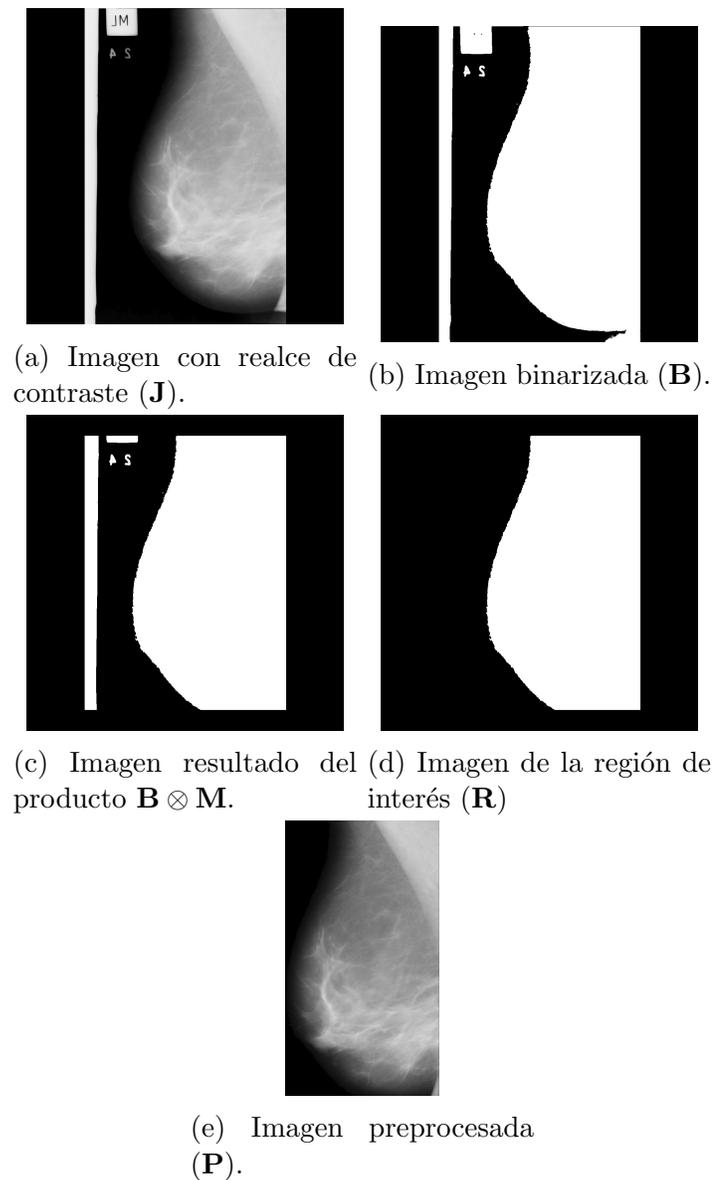


Figura 3: Pasos en la eliminación de artefactos

8.2. Remoción del musculo pectoral

Para eliminar la región correspondiente al músculo pectoral, se implementa una metodología de cuatro pasos: el primer paso es homogeneizar la orientación de las imágenes,

es decir, conseguir que todas las imágenes estén orientadas a una misma dirección. Para eso se divide la imagen preprocesada \mathbf{P} haciendo un corte vertical resultando en dos imágenes \mathbf{P}_I y \mathbf{P}_D , cada una de dimensiones $h' \times \frac{w'}{2}$. En cada una de estas imágenes, se cuentan los pixeles distintos de cero. Si el número de pixeles distintos de cero es mayor en la porción \mathbf{P}_I , entonces la imagen está orientada a la izquierda; de otro modo la imagen está orientada a la derecha [42]. En este trabajo se homogeneizaron todas las imágenes con orientación hacia izquierda (\mathbf{P}_F) como podemos ver en la Figura 4a.

En el segundo paso de esta metodología se busca eliminar el fondo de la imagen. Por lo tanto, se deben eliminar todas aquellas estructuras diferentes del músculo pectoral. Esto se consigue creando una estimación del fondo $\mathbf{K} \in \mathcal{N}^{h' \times w'}$ para luego suprimirla de la imagen usando la siguiente operación $\mathbf{P}_F - \mathbf{K}$, dando origen a la imagen $\mathbf{H} \in \mathcal{N}^{h' \times w'}$. La estimación \mathbf{K} es un gradiente que va de derecha a izquierda con una orientación de 45° , tal como se muestra en la Figura 4b. Lo anterior se realiza ya que se se busca preservar el músculo pectoral, el cual se supone está ubicado en el sector izquierdo de la imagen. Se escoge como valor a difuminar la intensidad más alta de la imagen \mathbf{P}_F garantizando así la eliminación de las estructuras irrelevantes de más alta intensidad como se puede observar en la Figura 4c.

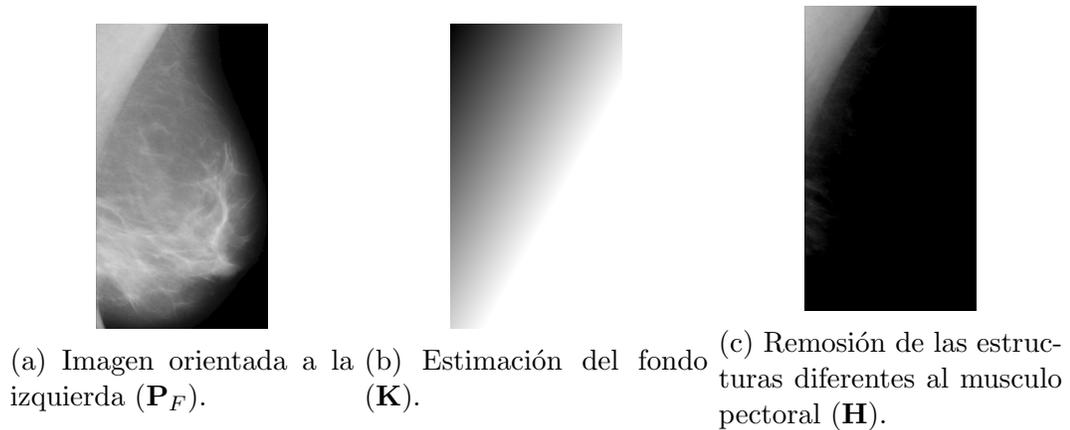


Figura 4: Pasos uno y dos de la remoción del musculo pectoral.

8.2.1. Segmentación del musculo pectoral con K-medias

En el paso tres usamos es algoritmo K-medias para separar el musculo pectoral de la región de interés (mama). La imagen \mathbf{H} esta compuesta por dos regiones (s_i), la

primera contiene el musculo pectoral y la segunda contiene el fondo de la imagen, este algoritmo de agrupamiento no supervisado es usado para clasificar los pixeles en dos clases (k) basándose en la distancia (\mathbf{d}) de las intensidades de cada uno de estos [43]. El algoritmo asume que existe un espacio de características (\mathbf{x}_j) en el cual los pixeles pueden ser agrupados alrededor de unos centroides ($\boldsymbol{\mu}_i$), los cuales son obtenidos minimizando la función objetivo:

$$\mathbf{d} = \min_{\mathbf{s}} \sum_{i=1}^k \sum_{\mathbf{x}_j \in s_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2 \quad (4)$$

En cada iteración (t) se actualizan los centroides imponiendo la condición:

$$\boldsymbol{\mu}_i^{(t+1)} = \frac{1}{|s_i^{(t)}|} \sum_{\mathbf{x}_j \in s_i^{(t)}} \mathbf{x}_j \quad (5)$$

El algoritmo converge cuando los centroides no cambien. Como resultado de lo anterior se obtiene la segmentación del músculo pectoral $\mathbf{H}_P \in \mathbf{H}$. Cabe anotar que en el paso anterior pueden surgir agujeros dentro de la región segmentada remanentes del algoritmo de agrupamiento K-medias. Por esta razón se rellenan todos estos agujeros y se aplica una operación morfológica de dilatación (\oplus) con un elemento estructurante cuadrado $\mathcal{S} \in \mathcal{N}^{m \times n}$ sobre \mathbf{H}_P , esto con el fin de reducir la distorsión y remover pixeles aislados [37].

8.2.2. Corrección del contorno con aproximación polinomial

El último paso de la metodología de remoción del músculo pectoral es el ajuste del contorno del propio músculo. Para esto se obtiene el perímetro de \mathbf{H}_P y se acondiciona antes de su corrección usando la operación morfológica de erosión (\ominus). El objetivo de la erosión es eliminar las líneas que se ajustan a los límites de la imagen y preservar solo el contorno que divide al músculo pectoral del resto del tejido mamario. Para esto se usa un elemento estructurante de línea cuya longitud debe ser muy menor en comparación con las líneas horizontales y verticales del contorno de la imagen (en este caso se ha escogido experimentalmente un elemento de longitud igual a 10 pixeles) en las orientaciones 90° y 0° . A continuación, se hace una aproximación del perímetro del músculo pectoral (ver Figura 6) utilizando un modelo polinomial de tercer grado sobre el borde de la imagen. Se elige una función polinomial cúbica debido a que la forma del músculo pectoral en la mayoría de los casos sigue una curvatura que puede ser mejor aproximada por una función de este orden [44].

Finalmente, se crea una máscara a partir del contorno corregido ($\mathbf{C} \in \mathcal{N}^{h' \times w'}$), tal como se muestra en la Figura 5b y se aplica la operación $\mathbf{C} \otimes \mathbf{P}_F$ para suprimir el músculo pectoral, obteniendo de esta manera la imagen mamográfica preprocesada ($\mathbf{L} \in \mathcal{N}^{h' \times w'}$) que se ve en la Figura 5c.

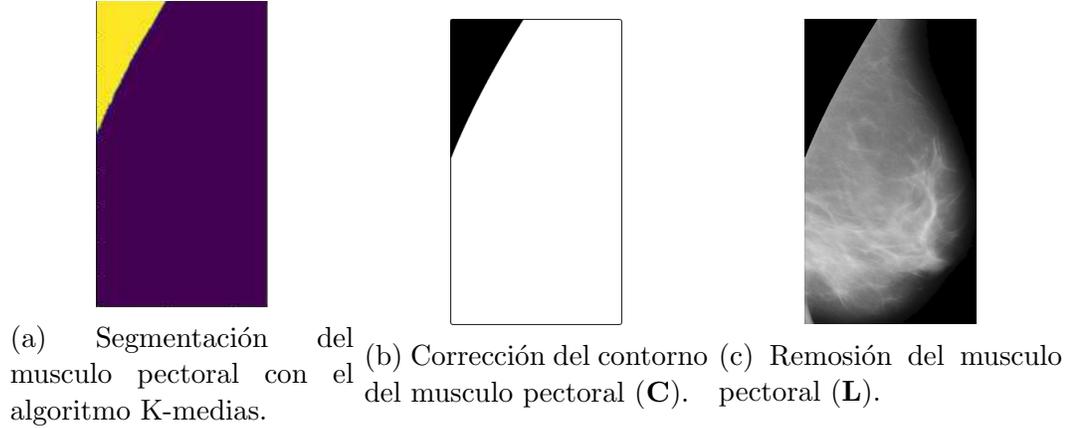


Figura 5: Pasos tres y cuatro para la remoción del músculo pectoral.

En este trabajo se realizaron aproximaciones con funciones de segundo grado y funciones de orden superior, y si bien estas funciones permiten una mayor flexibilidad, también se ven afectadas por la sensibilidad de los parámetros. En la Figura 6 puede verse la aproximación del contorno del músculo que se consiguió empleando una función cúbica.

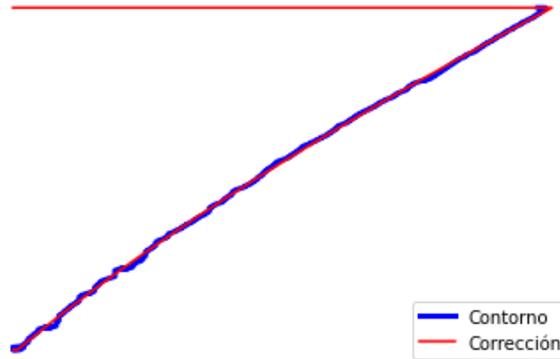


Figura 6: Aproximación polinomial del contorno del músculo pectoral.

8.3. Detección y localización de microcalcificaciones

La metodología propuesta para identificar y localizar MC consta de dos pasos generales. El primero es la identificación de una potencial región de interés ($\mathbf{W} \in \mathbf{L}$); en este trabajo la identificación de esta \mathbf{W} se realizó mediante una red neuronal convolucional (CNN), usada para clasificar si dicha región específica en la mamografía contiene una MC o no. Una vez que se ha identificado una \mathbf{W} que contiene MC, se aplica el segundo paso. La idea básica de este paso es mejorar el contraste de las MC para mejorar la precisión de la segmentación. Para esto, cada región se procesa en tres etapas: en la primera se elimina el fondo, en la segunda se eliminan las estructuras irrelevantes y en la tercera, se identifican las MC.

8.3.1. Detección de MC con CNN

Las CNN son un tipo de red neuronal artificial principalmente utilizada para la clasificación y reconocimiento de patrones en imágenes. Este tipo de red neuronal obtiene su nombre principalmente debido a que implementa la operación de convolución utilizando determinados kernels o filtros, los cuales se aplican a determinados datos de entrada. Esta implementación se da a través de la capa que caracteriza este tipo de red neuronal, la capa convolucional. En estas redes, los datos de entrada suelen pasar por todo un proceso de extracción de características, que comienza en la capa convolucional mencionada. Además, también es usual implementar capas que reduzcan la dimensionalidad de los datos y otras que realicen el proceso de clasificación según las características obtenidas. De esta forma, las capas que constituyen una red neuronal convolucional son las siguientes [45]:

- **Capa de convolución:** Esta capa permite la extracción de características a través de la aplicación de filtros o kernels ($\mathcal{F} \in \mathcal{M}^{m \times n}$) a los respectivos valores de entrada \mathbf{W} . Tales filtros estarán dados por los tensores de pesos, los cuales se van actualizando conforme se vaya realizando el entrenamiento de la red neuronal. Estos se aplican a determinadas regiones de los valores de la imagen recibidos como tensores, por lo que están conectados solo con un determinado campo receptivo local dado por el tamaño del filtro aplicado. Este proceso permite la obtención de características simples, las cuales pueden ser usadas para determinar otras más complejas en capas ocultas posteriores. Esta operación de convolución $\mathbf{W} \odot \mathcal{F}$, resulta en un valor que se conoce como un mapa de características, que será una matriz 2-dimensional que nos resaltará las áreas de la imagen que son más acordes con los valores del kernel.

- Capa de agrupamiento:** Esta capa permite reducir el tamaño espacial de la imagen de entrada, lo que conlleva a disminuir la cantidad de parámetros, la carga computacional y la memoria que se usa, para controlar el riesgo de sobreajuste (overfitting). Además, con esta reducción de la entrada de esta capa, se logra que la red neuronal tenga un aprendizaje y tolere pequeños cambios en la imagen. Sin embargo, esta capa tiene como característica que no cuenta con pesos, ya que solo ejecuta una operación de submuestreo $f_{X,Y}(\mathcal{S}) = \max_{a,b=0}^1 \mathcal{S}_{2X+a,2Y+b}$.
- Perceptrón multicapa:** Esta capa tiene la característica, en contraste con las anteriores, de que esta totalmente conectada con todos los valores de entrada. La salida (Y_j) de esta capa esta determinada por la multiplicación matricial usual de los pesos (W_{ij}) con los valores de entrada (X_i) y la suma de un sesgo (b_j) si se desea $Y_j = \sum_i^n W_{ij}X_i + b_j$. Esta capa es la de uso regular en cualquier tipo de arquitectura de red neuronal artificial. Los datos de salida de esta capa son suministrados a una función de costo que ha sido optimizada durante el entrenamiento y permite la clasificación de las imágenes de entrada de la red neuronal según las características encontradas.

Así pues, una red neuronal convolucional se construye cuando hay alguna combinación de estos tres tipos de capas. La arquitectura básica utilizada para una tarea de clasificación de imágenes se muestra en la Figura 7. Los parámetros de la red y la configuración del entrenamiento de la misma se especificaran en secciones posteriores.

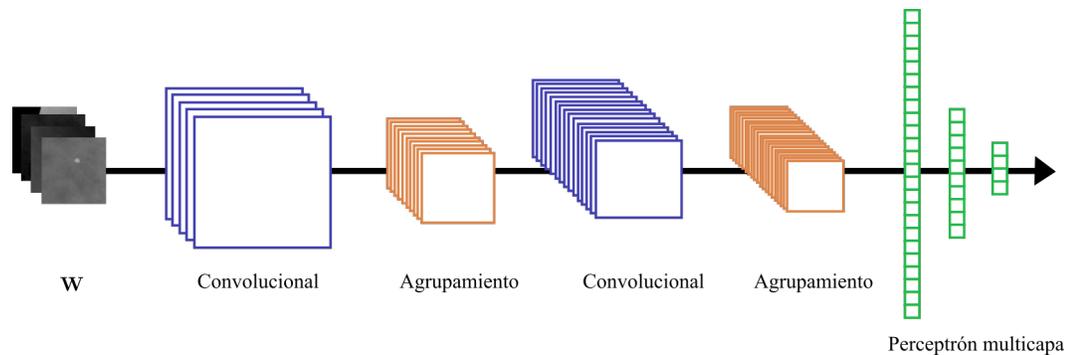
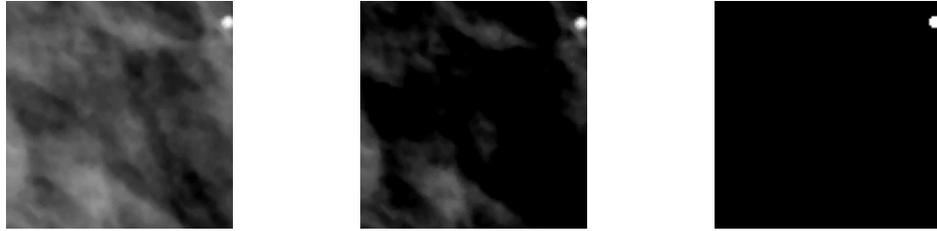


Figura 7: Arquitectura básica de una CNN.

8.3.2. Realce de contraste, segmentación y filtrado de MC

Finalmente en el segundo paso realizamos el contraste de las MC, esto se consigue removiendo el fondo de la imagen \mathbf{W} donde previamente fueron detectadas posibles MC, mediante la operación $\mathbf{M} = \mathbf{W} - \mathbf{W}_F$ (Figura 8b), donde \mathbf{W}_F representa el fondo de la imagen y cada uno de sus pixeles tiene como valor la intensidad promedio de \mathbf{W} (Figura 8a), posteriormente binarizamos la imagen resultante, para esto usamos nuevamente el algoritmo de umbralización de Otsu discutido previamente, obteniendo así el número de objetos (MC) contenidos en \mathbf{M} (Figura 8c), de la imagen binaria resultante se extrae el centroide de todos los objetos, estas serán las coordenadas relativas de las MC en la imagen, como se puede observar en el resultado final mostrado en la Figura 9, donde las coordenadas de las MC dentro de \mathbf{M} se han convertido a coordenadas globales de la imagen \mathbf{L} sumando el desplazamiento horizontal y vertical realizado en el muestreo.



(a) Región de interés con posible MC detectada (\mathbf{W}). (b) Realce de contraste (\mathbf{M}). (c) Binarización de la imagen \mathbf{M} .

Figura 8: Pasos para localización de MC.

Con el objetivo de eliminar estructuras irrelevantes que favorezcan la aparición de FP, se usa el hecho de que el diámetro de las MC varía entre 0,1mm y 1 mm lo que nos permite filtrar las estructuras clasificadas erróneamente como MC en el paso anterior. Esto significa que los objetos de forma lineal y mayores de 1mm se descartaron en función de la longitud de su eje mayor. Además, las MC tienden a aparecer como pequeños puntos brillantes circulares; por lo tanto, una intensidad basada en la característica se utilizó para distinguir los MC verdaderos. En resumen, los candidatos de MC que satisfacen uno o más de las siguientes condiciones fueron eliminados:

$$\begin{aligned}
 l_{obj} &> l_{max} \\
 A_{obj} &> A_{max} \\
 A_{obj} &< A_{min} \\
 I_{obj} &< I_{min}
 \end{aligned}$$

Donde l_{max} , A_{max} , A_{min} y I_{min} denotan la longitud máxima del semieje mayor, el área máxima, el área mínima y la intensidad promedio del objeto seleccionado, respectivamente. El valor de I_{max} es definido como $\lceil \frac{1}{res} \rceil$, donde res es la resolución de la imagen en milímetros; A_{max} es definida como $\lceil \pi (\frac{l_{max}}{2})^2 \rceil$; $A_{min} = 3$ y $I_{min} = 190$ fueron seleccionados experimentalmente.

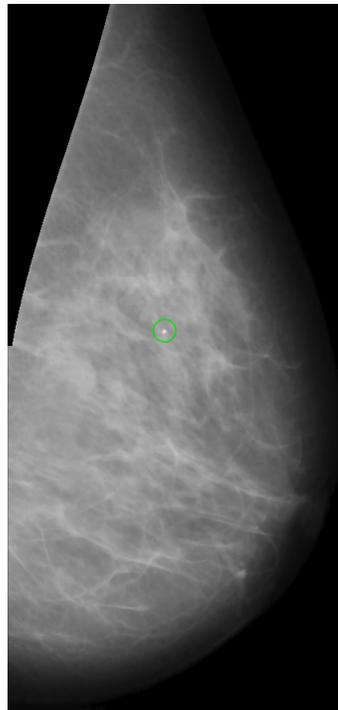


Figura 9: Resultado final del procesamiento, donde se muestra una mamografía sin artefactos, sin musculo pectoral y con las MC localizadas.

8.4. Clasificación de microcalcificaciones según su categoría BI-RADS

En esta sección se presenta una metodología para la recuperación de imágenes mediante un modelo CBIR el cual es apoyado por un sistema de clasificación de MC según la escala BI-RADS, para redireccionar la búsqueda en el proceso de emparejamiento. la metodología se divide en tres etapas, la primera consiste en la extracción de un total de 93 características, la segunda etapa es la clasificación, en la cual se implementa un ensamble de RNA y la tercera consiste en un sistema de búsqueda basado en la clase predicha en la etapa anterior y el emparejamiento basado en la distancia euclidiana.

8.4.1. Escala BI-RADS

Para estandarizar el análisis, la evaluación y el manejo de hallazgos mamográficos en este tipo de imágenes, el Colegio Americano de Radiología desarrolló una terminología estandarizada denominada Breast Imaging Reporting and Data System (BI-RADS). Esta escala estándar incluye además categorías para los hallazgos, los cuales son clasificados de acuerdo con su probabilidad de malignidad. Según el estándar BI-RADS hay siete categorías (de la 0 a la 6) para clasificar una mamografía [12]. La categoría BI-RADS 0 se da a imágenes en las que la evaluación es incompleta y son necesarios nuevos exámenes diagnósticos, bien sean nuevas tomas de imágenes mamográficas bajo otras condiciones o imágenes de ultrasonido. La categoría BI-RADS 1 se da a imágenes con evaluación negativa en las que no se han encontrado masas, calcificaciones sospechosas o áreas con distorsiones. La categoría BI-RADS 2 se da a imágenes consistentes con hallazgos benignos. La categoría BI-RADS 3 se otorga a imágenes con hallazgos probablemente benignos, en donde el riesgo de malignidad se ubica por debajo del 2%. Existen criterios muy estrictos para clasificar una imagen dentro de esta categoría y es quizás una de las categorías más complejas de clasificar por parte de los especialistas. La categoría BI-RADS 4 se otorga a imágenes en donde hay hallazgos con anormalidades sospechosas. Esta categoría se subdivide a su vez en subcategorías a, b y c. La subcategoría 4.a se da a imágenes en las que hay hallazgos con probabilidades de malignidad que oscilan entre el 2% y el 10%. La subcategoría 4.b tiene un rango intermedio de probabilidad de malignidad que va del 10% al 50%. Y finalmente, la subcategoría 4.c tiene un rango de probabilidad de malignidad del 50% al 95%. La categoría BI-RADS 5 se da a imágenes con hallazgos con una muy alta probabilidad de malignidad (superior al 95%). Y por último, la categoría BI-RADS 6 se da a imágenes en las que hay hallazgos patológicamente confirmados como posi-

tivos para malignidad.

8.4.2. Extracción de características

Radiomics es un método para el procesamiento de imágenes que ha tenido un gran auge en los últimos años, empleado con éxito en el campo de la oncología, incluido el carcinoma de ovario y especialmente aplicado en el diagnóstico del cáncer de seno, efectivo para el análisis cuantitativo y la predicción, usando grandes cantidades de imágenes médicas (Big data), y extrayendo información relevante sobre el tejido mamario a través de herramientas computacionales avanzadas (Aprendizaje automático) [16]. Para la etapa de extracción de características se utilizó **Pyradiomics** (librería de código abierto de Python) para extraer estadísticas de primer orden, basadas en la forma, la matriz de coocurrencia de niveles de gris, de longitud de niveles de gris, de zona y tamaño de niveles de gris, de diferencia de tonos de gris y de dependencia de niveles de gris, con 19, 10, 13, 16, 16, 5 y 14 características respectivamente [46].

8.4.3. Redes Neuronales Artificiales

Las RNA son una poderosa herramienta en el área de la clasificación y el reconocimiento de patrones. Nuestro principal interés en las RNA es el diagnóstico médico, concretamente el reconocimiento de patrones para mamografías digitales aplicado al problema de clasificación BI-RADS. El procesamiento estándar de las RNA se inicia tomando un vector de características $\mathbf{x} \in \mathcal{R}^N$ en la capa de entrada, cada valor (x_j) se propaga a lo largo de todas las capas ocultas, activando las unidades que se encuentran conectadas entre sí a la capa de salida (ver Figura 10), representada por la expresión $y(\mathbf{x}, \mathbf{w})_k = h(\sum_i w_{kj}x_j)$. Donde los w_{kj} son los pesos de las conexiones de las unidades y el $h(\cdot)$ es una función de activación que limita el valor de la amplitud de la salida de las unidades [47]. Los pesos (\mathbf{w}) se actualizan mediante un proceso de entrenamiento el cual consiste en minimizar una función de error $E(\mathbf{w})$, cada predicción de salida se contrasta con un valor validado previamente definido (etiqueta) esta información se retropropaga a través de la RNA, modificando los pesos para una predicción más precisa. La ecuación de actualización de los pesos se define como $\mathbf{w}^{\tau+1} = \mathbf{w}^{\tau} - \eta \nabla E_n(\mathbf{w}^{\tau})$ [48].

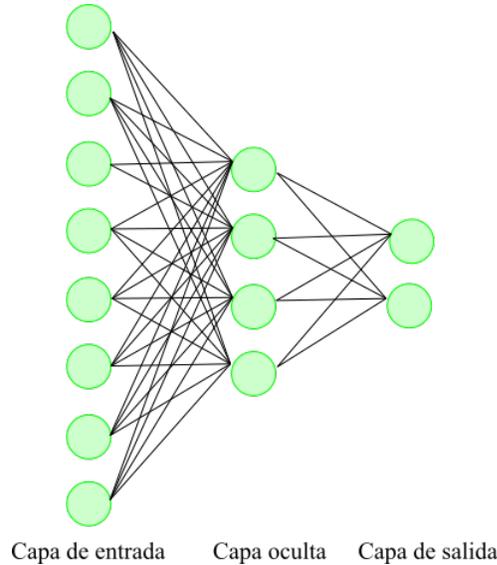


Figura 10: Arquitectura básica de una RNA.

Por otro lado para predecir cada categoría BI-RADS se diseñó un enfoque biclase, donde se le asigna un modelo de RNA a cada categoría de forma independiente (ensamble), por lo que para obtener un resultado global utilizamos la clase predicha por la mayoría de modelos, que tendrá una probabilidad de clase $p_c = \frac{n_c}{3}$, donde n_c es el número de veces que se predijo la categoría c para un modelo independiente, el factor de normalización $\frac{1}{3}$ corresponde al número de modelos que pueden predecir la clase c respectivamente. La categoría BI-RADS predicha se utilizará posteriormente como un soporte para la búsqueda de imágenes similares por parte del sistema CBIR.

8.4.4. Sistema de recuperación de imágenes de microcalcificaciones

Los modelos CBIR en el contexto de los sistemas DAO tienen como principal objetivo entregar a un especialista una serie de imágenes similares a una imagen consulta [31]. La estrategia de emparejamiento utilizada para nuestro CBIR se basa en encontrar las imágenes mejor emparejadas (más similares) del conjunto de datos, con respecto a una imagen de consulta \mathbf{q} . Consideremos el vector de características $\mathbf{x}_c \in \mathcal{X}$ que representa una ROI embebida en el espacio de características \mathcal{X} , donde cada vector de características está etiquetado con una clase c de acuerdo con la categoría BI-RADS determinada por un experto. Este es nuestro conjunto de datos, del que recuperaremos las imágenes más similares a \mathbf{q} . Para obtener un resultado más preciso en la búsqueda proponemos un soporte auxiliar de nuestro modelo de clasificación RNA,

asumiendo un espacio de características \mathcal{X} segmentado y la salida RNA (c) como una ruta de búsqueda, forzando al sistema CBIR a encontrar las imágenes más similares sólo en la sección $\mathcal{X}_c \subset \mathcal{X}$ predicha por el modelo, como se puede ver en la Figura 11.

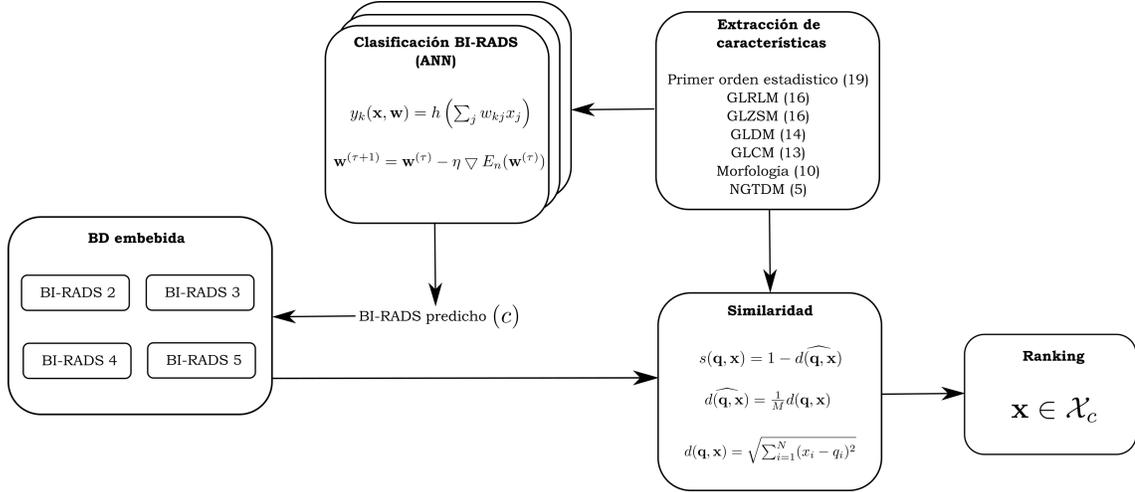


Figura 11: Esquema metodológico del CBIR.

Finalmente, para calcular la similitud (s) entre cada ROI ($\mathbf{x} \in \mathcal{X}_c$) e imagen consulta \mathbf{q} , se usó la expresión $s(\mathbf{q}, \mathbf{x}) = 1 - d(\widehat{\mathbf{q}}, \mathbf{x})$, donde $d(\widehat{\mathbf{q}}, \mathbf{x}) = \frac{1}{M}d(\mathbf{q}, \mathbf{x})$ es la distancia Euclidiana normalizada, y M es el valor de la distancia máxima medida en el espacio de referencia \mathcal{X} . Así, cuando la similitud $s \in [0, 1]$ toma el valor de 1 presenta una imagen muy similar y 0 representa una imagen no similar. Todas las imágenes se clasifican de la más a menos similares, y el sistema CBIR recupera las seis imágenes mejor posicionadas.

9. Marco experimental

En esta sección se describe la configuración de las pruebas realizadas para la validación del sistema construido a partir de los módulos que implementan las metodologías de preprocesamiento de mamografías y localización de clusters de MC establecidas previamente. En los experimentos se contempla el uso de dos bases de datos (BD) publicas mini-MIAS y Digital Database for Screening Mammography (DDSM), cuyas características se exponen en detalle. La validación estadística del sistema DAO se hará en función de métricas de rendimiento supervisadas, entre las que se encuentran: acierto de los sistemas de clasificación, medidas de sensibilidad y especificidad, tasa de falsos positivos, falsos negativos y coeficiente Dice para evaluar traslape entre contornos reales y predichos, entre otros.

Todos los algoritmos aplicados para las pruebas a continuación descritas fueron elaborados sobre el lenguaje de programación Python y la implementación de los métodos propuestos se pueden encontrar libremente en un repositorio en línea ¹.

9.1. Bases de datos

Este proyecto de investigación se clasificó como una investigación sin riesgo, ya que no se realizó ninguna intervención o modificación intencionada de variables biológicas, fisiológicas, psicológicas o sociales de las personas que participan en el estudio. Se contó con dos BD públicas de mamografías. Por una parte, se usaron imágenes de acceso libre mini-MIAS, la cual contiene 322 imágenes, de los cuales 28 casos corresponden a MC sospechosas respectivamente, confirmadas por biopsias (Gold Standard) [49]. Además, se usaron los datos de la BD DDSM la cual contiene 1511 imágenes, todas con MC confirmadas por biopsias [50], ambas BD contienen imágenes es proyección craneocaudal (CC) y MLO. Los pacientes, cuyas imágenes fueron usadas, firmaron un consentimiento informado. Los datos fueron anonimizados para asegurar la privacidad de los pacientes.

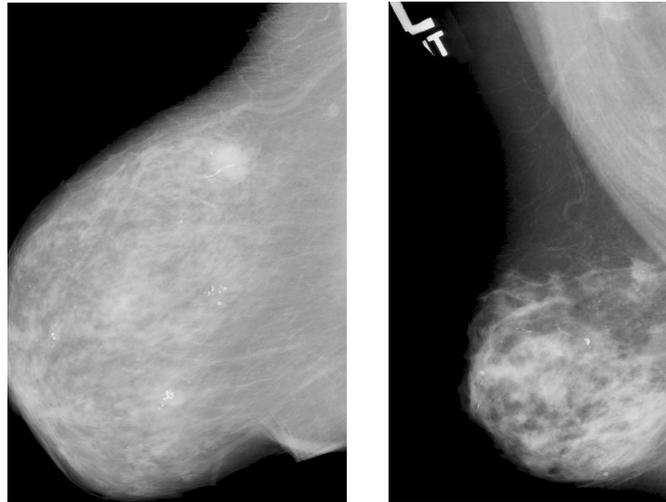
BD	Numero de imágenes	Tamaño(pixeles)	Formato	Resolución
mini-MIAS	322	1024 × 1024	.pgm	200 μ /p
DDSM	1511	~ 3016 × 4616	.dcm	50 μ /p

Cuadro 1: Especificaciones de las BD.

¹<https://github.com/KevinGomez26/codigos-MC/blob/main/algoritmos.py>

9.2. Resultados de las pruebas de eliminación de artefactos

El cuadro 2 muestra los resultados obtenidos para la eliminación de ruido y supresión de artefactos aplicados sobre DDSM y mini-MIAS, comparados con otros métodos usados en el estado del arte. La métrica reportada indica el nivel de acierto (ACC) en la segmentación de la mama comprobada a través de una inspección visual, similar a los trabajos reportados en la literatura [51, 38, 52, 53] en donde un radiólogo evalúa los resultados visualmente, determinando el número de imágenes segmentadas de forma correcta (ISC), posteriormente se obtiene el valor numérico de la métrica mediante la expresión $ACC = \frac{ISC}{N} 100\%$, donde N es el número de imágenes totales.



(a) Mama segmentada de forma correcta. (b) Mama segmentada de forma incorrecta.

Figura 12: Ejemplos segmentación correcta e incorrecta.

El radiólogo encargado identifica que la imagen resultado no contenga ningún tipo de artefacto o ruido y que el contorno de la mama haya sido segmentado correctamente. En la Figura 12 se muestra una imagen que ha sido segmentada correctamente contrastada con una imagen que no ha sido segmentada de forma correcta por el método implementado en este trabajo.

Propuesto por	DDSM	mini-MIAS
Nosotros	99.87 %	99.69 %
Qayyum [51]	-	99.37 %
Slavkovic [38]	-	97.51 %
Yoon [52]	-	93.16 %
Wei [53]	94.90 %	-

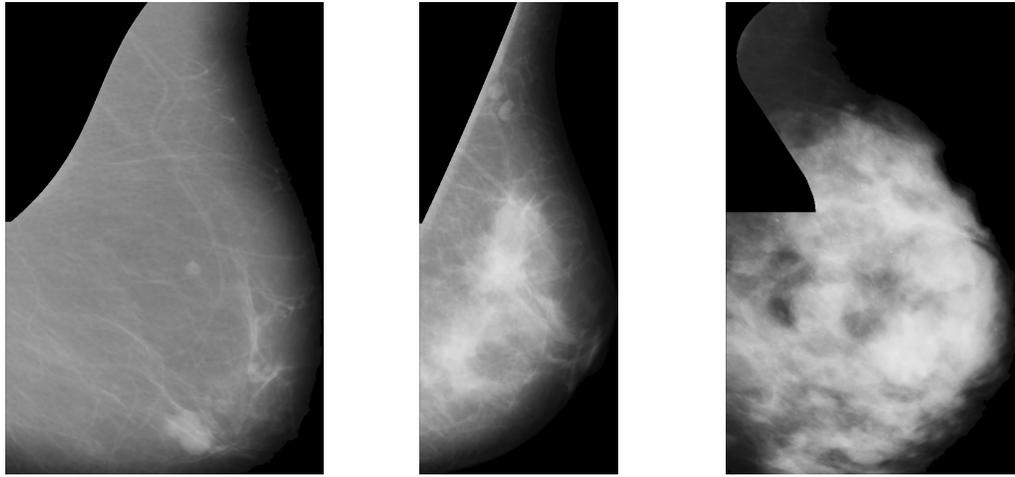
Cuadro 2: Resultados de las pruebas de eliminación de artefactos comparados con otros métodos propuestos en la literatura.

9.3. Resultados de las pruebas de remoción del musculo pectoral

Para la validación de este experimento se usaron todas las imágenes presentes en la base de datos mini-MIAS y 794 imagines de la base de datos DDSM correspondientes solo a las proyecciones MLO (debido a que este es el conjunto de imágenes en el que se encuentra presente el musculo pectoral). La métrica reportada indica el nivel de acierto en la segmentación y remoción del musculo pectoral comprobada a través de una inspección visual basada en tres categorías [25]:

- **Excelente (E):** El contorno del musculo pectoral detectado emula exactamente el contorno real y no es posible identificar tipo de desviación de forma visual, ademas la región pectoral detectada coincide con la zona pectoral en cuestión y esta no se solapa de forma abrupta con el tejido mamario.
- **Aceptable (A):** La región detectada especifica correctamente el musculo pectoral y su contorno es aceptable para el propósito específico.
- **Inaceptable (I):** La región detectada no es la adecuada debido a su enorme solapamiento con el tejido mamario y es inaceptable para el propósito específico.

En la Figura 13 puede apreciarse una comparativa de estas tres categorías, donde la Figura 13b representa una imagen donde se ha removido satisfactoriamente el musculo pectoral y el radiólogo a determinado que es un resultado excelente, en la Figura 13b se determino que la remoción ha sido aceptable pese a que el contorno real y el contorno detectado no coincidan exactamente, y finalmente en la Figura 13c se ha determinado que el resultado el inaceptable ya que la región detectada se solapa de forma notable con el tejido mamario.



(a) Musculo pectoral removido de forma excelente. (b) Musculo pectoral removido de forma aceptable. (c) Musculo pectoral removido de forma inaceptable.

Figura 13: Ejemplos de remoción del musculo pectoral.

Los resultados obtenidos para la la BD mini-Mias fueron 74,54 %, 15,04 % y 10,42 % correspondientes a las categorías E, A e I respectivamente. Para la BD DDSM fueron 73,80 %, 20,91 % y 5,29 % correspondientes a las categorías E, A e I respectivamente. En el Cuadro 3 se muestran los resultados de acierto, en el cual se tienen en cuenta tanto las imágenes con resultados excelentes como aceptables $ACC = E + A$, contrastado con otros resultados obtenidos en la literatura.

Propuesto por	BD(imágenes)	Acierto
Nosotros	mini-MIAS	89.58 %
	DDSM(794)	94.71 %
Shinde [20]	mini-MIAS	93.70 %
Qayyum [51]	mini-MIAS	93.00 %
Slavkovic [38]	mini-MIAS	87.57 %
Sreedevi [54]	mini-MIAS(161)	90.06 %
Raba [55]	mini-MIAS	86.00 %
Saltanat [56]	mini-MIAS	92.00 %
Wang [57]	DDSM(80)	91.00 %
Alam [43]	mini-MIAS	90.30 %
Li [58]	DDSM(100)	92.00 %

Cuadro 3: Resultados del método de remoción del musculo pectoral comparados con otros resultados reportados en la literatura.

9.4. Resultados de las pruebas de detección y localización de MC

Para entrenar el modelo de detección de MC se uso una BD construida a mano a partir de la extracción de regiones de interés (ROI) de tamaño 100×100 pixeles, este tamaño es el adecuado para la identificación de MC sin que se vea afectada la imagen por distorsión geométrica causada por la pixelación [5]. Las ROI podrían contener o no MC, en total 2014 imágenes fueron extraídas y etiquetadas como no-MC (si no contenían MC) y como si-MC (si contenían MC). Para el entrenamiento se uso una partición del 70 % (1409 imágenes) de la BD y 30 % (605 imágenes) restante para la validación del mismo, los resultados obtenidos se pueden observar en el Figura 14, la cual muestra la matriz de confusión.

Por otro lado la arquitectura de la CNN esta compuesta por una capa de entrada de tamaño 100×100 ; cuatro capas de convolución con filtros de tamaño 3×3 , y una cantidad de 16, 32, 64 y 128 filtros respectivamente; entre cada capa convolucional se adiciono una capa de agrupamiento (maxpooling) de tamaño 2×2 para extraer características que se puedan ver afectadas por pequeñas variaciones en la imagen [59], finalmente se acoplo un perceptrón multicapa con 128 unidades de entrada con funciones de activación de rectificado lineal para un rápido procesamiento en la etapa de entrenamiento [60], ademas de 2 unidades de salida con funciones de activación softmax, de esta forma cada salida de la red se puede interpretar como la probabilidad de pertenencia a cada clase [61]. Ademas se uso un optimizador Adam (Adaptative Moment Estimation) para acelerar la búsqueda del punto estacionario, ya que a diferencia de otras técnicas, la tasa de entrenamiento se adapta en función de las magnitudes de los pesos, de esta manera el algoritmo converge rápidamente y las direcciones no tan comunes que tome el gradiente son canceladas [48].

Predicción	no-MC	297 95%	1 0%
	si-MC	17 5%	290 100%
		no-MC	si-MC
		Actual	

Figura 14: Matriz de confusión resultante de la etapa de entrenamiento.

Por otro lado las épocas fueron ajustados a el valor de 30. La arquitectura y los parámetros de la CNN fueron seleccionados experimentalmente.

En el Cuadro 4 se pueden observar los resultados de la localización de MC donde se usaron 287 imágenes de la BD mini-MIAS y 692 imágenes de la BD DDSM (correspondientes a los mejores resultados obtenidos por el sistema). Las métricas reportadas son, la sensibilidad (s) la cual se define como $s = \frac{TP}{TP+FN}$, donde TP y FN son las MC clasificadas de forma correcta y las MC clasificadas de forma incorrecta respectivamente; mientras que por otro lado también se reporto la tasa de FP por cada imagen (r) la cual se define como $r = \frac{FP}{N}$, donde N es el numero de imágenes.

Propuesto por	DB(imágenes)	Sensibilidad	FP por imagen
Nosotros	mini-MIAS(287)	77 %	0.79
	DDSM(692)	81 %	0.16
Yang [62]	Privada(292)	85 %	0.40
Wang [26]	Privada(292)	90 %	0.69
Liu [63]	Privada(205)	92 %	2.30
El-Naqa [64]	Privada(76)	94 %	1.00
Gallardo [65]	DDSM(100)	82 %	2.55

Cuadro 4: Sensibilidad y tasa de FP por imagen comparados con otros resultados reportados en la literatura.

9.5. Resultados de las pruebas de clasificación de MC según su categoría BI-RADS

En la Figura 15 se presenta la matriz de confusión relacionada a los resultados de clasificación del modelo de ensamble de RNA, dicho modelo fue entrenado con la BD DDSM, la cual consiste en 1,872 ROI etiquetadas con las categoría BI-RADS (0-5), la categoría Bi-RADS 0 fue excluida del análisis ya que esta representa imágenes que aun no tienen una valoración dada por un especialista. Previo a la etapa de entrenamiento se llevo acabo un aumento de datos, creando imágenes a partir de las ya existentes usando rotaciones de entre 10° – 70° lo que incremento el tamaño de la BD a 4,171 imágenes. Para el entrenamiento del modelo se uso el 80 % (3,337 imágenes) de la BD y el 20 % (834 imágenes) para la validación, un total de 3,000 iteraciones, y el Gradiente Descendente Estocástico con Momento (SGDM), con un factor de aprendizaje de 0,005, y un momento de 0,9.

Actual	BI-RADS 2	958 87.2%	8 0.8%	110 10.0%	22 2.0%
	BI-RADS 3	57 5.4%	915 86.6%	70 6.6%	14 1.4%
	BI-RADS 4	153 16.3%	53 5.6%	706 75.2%	27 2.9%
	BI-RADS 5	71 7.1%	45 4.5%	113 11.2%	772 77.2%
		BI-RADS 2	BI-RADS 3	BI-RADS 4	BI-RADS 5
		Predicha			

Figura 15: Matriz de confusión modelo de ensamble de RNA.

Por otro lado la arquitectura implementada para cada modelo de RNA consto de 3 capas ocultas con 32, 16 y 8 unidades respectivamente, con funciones de activación ReLU, y 2 unidades de en la capa de salida con función de activación softmax, determinados de forma experimental.

En el Cuadro 5 se muestra el nivel de acierto (a) general en la clasificación, el cual se obtuvo por medio de la expresión $a = \frac{TP+TN}{N}$ obteniendo un valor de 0,8185 para la tarea multiclase (4 categorías) en comparación con otros enfoques para diferentes objetivos de clasificación como la tarea de clasificación BI-RADS 4 y la tarea BI-RADS 2 vs 3 vs 4 vs ninguna lesión (4 categorías).

Propuesto por	Tarea	Acierto
Nosotros	BI-RADS (2,3,4,5)	82 %
Avalos et al. [16]	BI-RADS (2,3,4) y no lesionado	71 %
Lei et al. [18]	Bi-RADS 4	80 %

Cuadro 5: Resultados de la clasificación de MC según la escala BI-RADS comparados con otros resultados reportados en la literatura.

9.6. Diseño de la interfaz del sistema DAO

Todas las etapas de procesamiento descritas en las secciones anteriores fueron programadas en el software **Python** en la versión 3.8, para el diseño de la interfaz gráfica de usuario (GUI) del sistema DAO se optó por usar la librería de uso libre **tkinter**, esta librería proporciona herramientas para crear GUI sofisticadas, suministrando diversos componentes (botones, sliders, paneles, menús deslizantes, visualizador de imágenes, etc.) que pueden ser utilizados en combinación para producir prácticamente cualquier diseño de la GUI. A continuación se presenta, en la Figura 16, la interfaz diseñada para el sistema DAO.

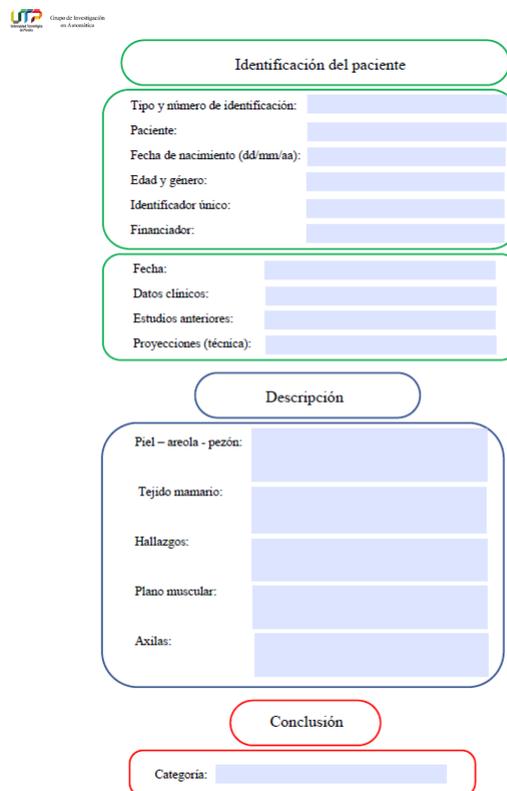


Figura 16: Panel inicial del sistema DAO.

El panel inicial está compuesto por dos secciones principales: la primera es la barra de botones, en la cual se encuentran los botones (ocho) que permiten ejecutar diversas acciones sobre la mamografía bajo estudio, la segunda sección es el panel donde se visualiza la mamografía que se desea analizar. El botón **cargar imagen** permite abrir un explorador de archivos donde el usuario podrá seleccionar la mamografía a analizar. El botón **eliminar artefactos** realiza las operaciones de preprocesamiento, eliminando el ruido y los artefactos contenidos en la mamografía, además de segmentar el seno y reducir la imagen a la región de interés. El botón **remover pectoral** remueve el músculo pectoral de la mamografía (asegurarse de que la imagen sea tomada en proyección MLO). El botón **localizar MC** permite al usuario localizar de forma automática las MC contenidas en la mamografía. El botón **guardar** abre un ex-

plorador de archivos que le permite al usuario guardar la imagen que desee. El botón **ayuda** mostrara un manual con toda la información necesaria para que el usuario pueda familiarizarse con el software.

Por otro lado el botón **información** abre un archivo .pdf como se puede ver en la Figura 17 donde el usuario podrá visualizar y editar los datos asociados al paciente en cuestión.



UTP Centro de Investigación en Alimentos

Identificación del paciente

Tipo y número de identificación:

Paciente:

Fecha de nacimiento (dd/mm/aa):

Edad y género:

Identificador único:

Financiador:

Fecha:

Datos clínicos:

Estudios anteriores:

Proyecciones (técnica):

Descripción

Piel – areola - pezón:

Tejido mamario:

Hallazgos:

Plano muscular:

Axilas:

Conclusión

Categoría:

Figura 17: Panel de información del paciente.

Este panel reúne toda la información asociada al paciente en tres campos diferentes: el primero es la identificación del paciente, aquí se encuentra el numero de identificación, el nombre, la edad, el genero y algunos datos relacionados con el examen medico,

como la fecha y proyecciones de la mamografía. El segundo campo es la Descripción, esta información la digita un especialista el cual se encarga de identificar las diferentes características asociadas a los tejidos y otros componentes de la mamografía, así como indicar los hallazgos encontrados. Finalmente el tercer campo es la conclusión, aquí se indica la categoría Bi-rads asociada a la mamografía determinada por un especialista.

Finalmente el botón **buscar similares** abre el panel de recuperación de información como se puede observar en la Figura 18 En este panel el usuario podrá visualizar el conjunto de imágenes con MC obtenida en etapas de procesamiento anteriores, adicional mente el sistema entregara un diagnostico referente a la categoría Bi-rads de la imagen MC en cuestión, además el sistema también arrojará seis imágenes consideradas similares posicionadas en un ranking dado por el nivel de similitud encontrado teniendo como referente la imagen consulta.

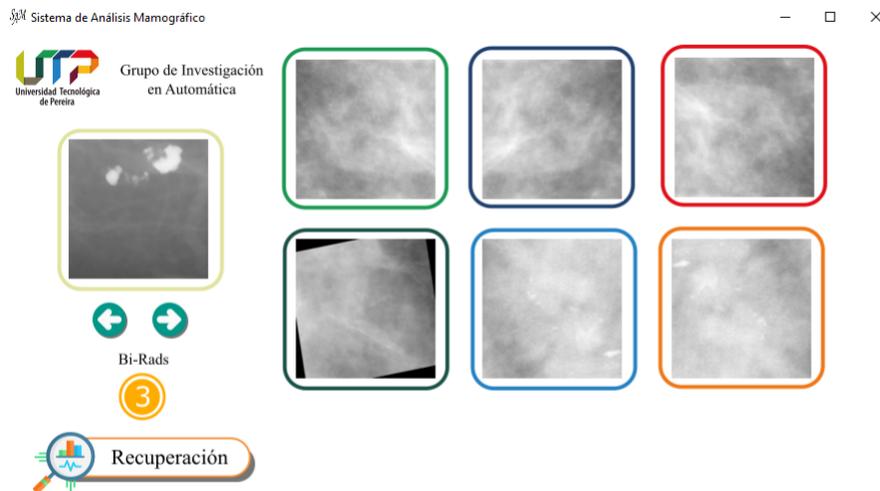


Figura 18: Panel de recuperación de información.

10. Conclusiones

En este trabajo se presenta una metodología alternativa para el preprocesamiento y localización de MC en mamografías crudas. En consecuencia, se genera un esquema metodológico que posteriormente permitió la implementación de un sistema DAO para el apoyo en la detección temprana del cáncer de mama. Inicialmente se realizó el preprocesamiento, la cual cuenta con dos etapas, la primera es la eliminación de artefactos, para cual se usó el método de realce de contraste AGCWD para incrementar la intensidad de los píxeles pertenecientes a la mama en relación a los del fondo de la imagen, además también se usó el método de umbralización de Otsu para segmentar la mama, y finalmente con el método de componentes conectados se desecharon los artefactos, con resultados que evidentemente mejoran los conseguidos en otros trabajos (ver Cuadro 2).

Por otro lado la segunda etapa consiste en la remoción del músculo pectoral, la cual se consiguió mediante la estimación y eliminación del fondo para remover las estructuras irrelevantes, en combinación con el algoritmo de segmentación K-medias usado para identificar el músculo pectoral, para refinar los resultados obtenidos después de la segmentación del músculo pectoral se propuso un método de ajuste del contorno del músculo por medio de una aproximación polinomial de orden tres, los resultados obtenidos mediante inspección visual (ver Cuadro 3) demuestran el éxito del método a la hora de segmentar el músculo pectoral, esto se debe en gran parte a que el algoritmo K-medias determina unos píxeles prototipos (μ) en la imagen, lo cual es muy conveniente debido a que se asume que el músculo pectoral tiene una intensidad homogénea por lo que solo los píxeles más parecidos al prototipo adecuado garantizan la correcta segmentación de dicha región.

También se implementó un módulo de detección y localización de MC. La detección se consiguió mediante el uso de una CNN entrenada para reconocer los patrones relevantes de una MC, esto se logró entrenando el modelo con una BD hecha manualmente, los resultados de dicho entrenamiento (ver Figura 14) fueron bastante satisfactorios, demostrando así que el modelo discriminaba correctamente una clase de la otra. Posteriormente cada mamografía fue submuestreada ($\mathbf{W} \in \mathbf{L}$) donde cada Roi fue clasificado como potencial Roi con MC o no. Finalmente si un Roi contenía MC pasaba a la etapa de localización, donde a cada imagen se le realizaba un realce de contraste eliminando las estructuras que aparecen en el fondo de la imagen, por medio de la sustracción del fondo, finalmente se filtraron la mayor cantidad posible de FP (estructuras clasificadas erróneamente como MC) asumiendo algunas carac-

terísticas geométricas y de intensidad de las MC, las cuales permiten descartar otro tipo de estructuras. Los resultados obtenidos de sensibilidad y tasa de FP por imagen (ver Cuadro 4) indican que esta metodología es competitiva en comparación a otros trabajos propuestos en la literatura.

Ademas se implemento una metodología que consta de dos etapas, la primera es la clasificación MC en la categoría BI-RADS (2, 3, 4, 5), problema de clasificación multi-clase, abordado como una tarea de clasificación binaria en un modelo de ensamble de RNA, para entrenar el modelo se utilizó la BD DDSM con un aumento de imágenes para remover el desbalance de datos. La extracción de características se implementó con el paquete Pyradiomics Python, se extrajeron un total de 93 características. La segunda etapa incorpora un proceso de emparejamiento, toma la salida del modelo de conjunto de RNA para apoyar la búsqueda a lo largo de todo el conjunto de datos para la recuperación de imágenes (ver Figura 11). El soporte se entiende como una ruta, esto permite que el proceso de emparejamiento busque sólo en una “sección” específica de la BD haciendo el emparejamiento más preciso debido a que el sistema CBIR sólo recupera imágenes similares de la misma categoría BI-RADS de la imagen consulta. Para ranquiar las imágenes se utilizó el valor de similitud entre ellas calculado a partir de la distancia euclidiana normalizada. En el Cuadro 5 se presento la evaluación en términos de la métrica a para el modelo de ensamble de RNA, los resultados fueron relativamente buenos para la clasificación MC, superando a los métodos del estado del arte.

Asimismo, se implemento un prototipo de software (ver Figura 16) el cual acopla todas las etapas nombradas anteriormente, con el añadido de que este sistema DAO cuenta con un modulo de visualización y edición de datos asociados al paciente (ver Figura 17) y un sistema de recomendación, el cual dictamina una categoría Bi-rads dado un Roi y entrega un conjunto de imágenes similares, para que un especialista pueda contrastar la imagen en cuestión con las entregadas por el sistema y así reforzar o considerar su opinión inicial.

Todos los métodos fueron validados en dos BD, mini-MIAS y DDSM, las cuales son ampliamente usadas en la literatura. Los resultados alcanzados en este trabajo muestran que, efectivamente, la metodología propuesta mejora el desempeño en la segmentación de la mama y musculo pectoral, así como la clasificación de MC, en comparación con los métodos propuestos en la literatura. Alentado por estos resultados prometedores, se propone como trabajo futuro, robustecer cada una de las etapas planteadas en este trabajo para aumentar la capacidad de discriminación de

los clasificadores y mejorar la exactitud de los segmentadores, además de incorporar módulos de detección y localización de masas, asimetrías y distorsiones arquitectónicas, biomarcadores claves en el diagnóstico temprano del cáncer de mama. Finalmente se pretende mejorar el sistema de recomendación con un circuito de retroalimentación que le permita al usuario dar una calificación al trabajo apoyo dado por el sistema DAO.

11. Resultados académicos

A continuación se presentan los artículos publicados y sometidos a congresos y revistas internacionales, así como proyectos de investigación y aplicativos software desarrollados durante la ejecución de este trabajo.

- Gómez, Kevin Alejandro Hernández, Julian D. Echeverry-Correa, and Álvaro Ángel Orozco Gutiérrez. “Automatic Pectoral Muscle Removal and Microcalcification Localization in Digital Mammograms.” *Healthcare Informatics Research* 27.3 (2021): 222-230.
- Hernández, Kevin Alejandro, D. Cárdenas Peña, and Álvaro A. Orozco. “A space-structure based dissimilarity measure for categorical data.” *International Journal of Electrical and Computer Engineering (IJECE)* 11.1 (2021): 620-627.
- Ariosto Serna, Luis, Kevin Alejandro Hernández, and Piedad Navarro González. “A k-means clustering algorithm: Using the chi-square as a distance.” *International Conference on Human Centered Computing*. Springer, Cham, 2018.
- Cardona, LA Serna, KA Hernández Gómez, and Álvaro A. Orozco-Gutierrez. “La Characterization of high school students in the department of Risaralda using the Chi-Square metric.” *Scientia et Technica* 26.2 (2021): 119-126.
- Computer-aided diagnosis based on microcalcification classification-driven image retrieval for breast cancer diagnosis (Sometido).
- Prototipo de sistema de apoyo al diagnostico medico para la detección temprana del cáncer de mama - Sistema de Análisis Mamográfico (SAM).

12. Agradecimientos

Manifiesto mi gratitud al grupo de investigación en Automática y a su director Álvaro Á. Orozco Gutiérrez, quienes pusieron a mi disposición las herramientas necesarias para el ejecución de este trabajo, el cual fue desarrollado en el marco del proyecto de investigación “Prototipo de un sistema de recuperación de información por contenido orientado a la clasificación de grupos de microcalcificaciones en mamografías - PROTOCAM”. con código 6-19-1, financiado por la vicerectoria de investigación de la Universidad Tecnológica de Pereira.

13. Bibliografía

- [1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, “Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.
- [2] X. Li, H. Gao, Z. Chen, L. Zhang, X. Zhu, S. Wang, and W. Peng, “Diagnosis of breast cancer based on microcalcifications using grating-based phase contrast ct,” *European radiology*, vol. 28, no. 9, pp. 3742–3750, 2018.
- [3] J. Hernández-Capistrán, J. F. Martínez-Carballido, and R. Rosas-Romero, “False positive reduction by an annular model as a set of few features for microcalcification detection to assist early diagnosis of breast cancer,” *Journal of medical systems*, vol. 42, no. 8, p. 134, 2018.
- [4] M. Moghbel, C. Y. Ooi, N. Ismail, Y. W. Hau, and N. Memari, “A review of breast boundary and pectoral muscle segmentation methods in computer-aided detection/diagnosis of breast mammography,” *Artificial Intelligence Review*, pp. 1–46, 2019.
- [5] T. Basile, A. Fanizzi, L. Losurdo, R. Bellotti, U. Bottigli, R. Dentamaro, V. Didonna, A. Fausto, R. Massafra, M. Moschetta *et al.*, “Microcalcification detection in full-field digital mammograms: A fully automated computer-aided system,” *Physica Medica*, vol. 64, pp. 1–9, 2019.
- [6] Breastcancer.org, “Diagnosis of dcis,” <https://www.breastcancer.org/symptoms/types/dcis/diagnosis>, 2018, *Accessed* : 2020 – 04 – 25.
- [7] A. C. S. (ACS), “What is breast cancer?” <https://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html>, 2019, *Accessed* : 2020–04–23.
- [8] M. El-Shenawee, “Electromagnetic imaging for breast cancer research,” in *2011 IEEE Topical Conference on Biomedical Wireless Technologies, Networks, and Sensing Systems*. IEEE, 2011, pp. 55–58.
- [9] A. M. Hassan and M. El-Shenawee, “Review of electromagnetic techniques for breast cancer detection,” *IEEE Reviews in Biomedical Engineering*, vol. 4, pp. 103–118, 2011.

-
- [10] W. H. O. (WHO), “Breast cancer,” <https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/>, 2020, *Accessed* : 2020 – 04 – 25.
- [11] V. A. Karale, T. Singh, A. Sadhu, N. Khandelwal, and S. Mukhopadhyay, “Reduction of false positives in the screening cad tool for microcalcification detection,” *Sādhanā*, vol. 45, no. 1, p. 44, 2020.
- [12] C. D’Orsi, L. Bassett, S. Feig *et al.*, “Breast imaging reporting and data system (bi-rads),” *Breast imaging atlas, 4th edn. American College of Radiology, Reston*, 2018.
- [13] S. Rahimeto, T. G. Debelee, D. Yohannes, and F. Schwenker, “Automatic pectoral muscle removal in mammograms,” *Evolving Systems*, pp. 1–8, 2019.
- [14] A. Rampun, K. López-Linares, P. J. Morrow, B. W. Scotney, H. Wang, I. G. Ocaña, G. Maclair, R. Zwiggelaar, M. A. G. Ballester, and I. Macía, “Breast pectoral muscle segmentation in mammograms using a modified holistically-nested edge detection network,” *Medical image analysis*, vol. 57, pp. 1–17, 2019.
- [15] M. Alsheh Ali, M. Eriksson, K. Czene, P. Hall, and K. Humphreys, “Detection of potential microcalcification clusters using multivendor for-presentation digital mammograms for short-term breast cancer risk estimation,” *Medical physics*, vol. 46, no. 4, pp. 1938–1946, 2019.
- [16] E. D. Avalos-Rivera and A. D. J. Pastrana-Palma, “Classifying microcalcifications on digital mammography using morphological descriptors and artificial neural network,” in *IEEE CACIDI 2016-IEEE Conference on Computer Sciences*. IEEE, 2016, pp. 1–4.
- [17] E. S. McDonald, A. M. McCarthy, S. P. Weinstein, M. D. Schnall, and E. F. Conant, “Bi-rads category 3 comparison: probably benign category after recall from screening before and after implementation of digital breast tomosynthesis,” *Radiology*, vol. 285, no. 3, pp. 778–787, 2017.
- [18] C. Lei, W. Wei, Z. Liu, Q. Xiong, C. Yang, M. Yang, L. Zhang, T. Zhu, X. Zhuang, C. Liu *et al.*, “Mammography-based radiomic analysis for predicting benign bi-rads category 4 calcifications,” *European journal of radiology*, vol. 121, p. 108711, 2019.
- [19] S. Aminololama-Shakeri, C. K. Abbey, P. Gazi, N. D. Prionas, A. Nosratieh, C.-S. Li, J. M. Boone, and K. K. Lindfors, “Differentiation of ductal carcinoma in-situ

- from benign micro-calcifications by dedicated breast computed tomography,” *European journal of radiology*, vol. 85, no. 1, pp. 297–303, 2016.
- [20] V. Shinde and B. T. Rao, “Novel approach to segment the pectoral muscle in the mammograms,” in *Cognitive Informatics and Soft Computing*. Springer, 2019, pp. 227–237.
- [21] R. Shen, K. Yan, F. Xiao, J. Chang, C. Jiang, and K. Zhou, “Automatic pectoral muscle region segmentation in mammograms using genetic algorithm and morphological selection,” *Journal of digital imaging*, vol. 31, no. 5, pp. 680–691, 2018.
- [22] G. Toz and P. Erdogmus, “A single sided edge marking method for detecting pectoral muscle in digital mammograms,” *Engineering, Technology & Applied Science Research*, vol. 8, no. 1, pp. 2367–2373, 2018.
- [23] M. S. Salama, A. S. Eltrass, and H. M. Elkamchouchi, “An improved approach for computer-aided diagnosis of breast cancer in digital mammography,” in *2018 IEEE international symposium on medical measurements and applications (MeMeA)*. IEEE, 2018, pp. 1–5.
- [24] İ. I. Eserner, S. Ergin, and T. Yüksel, “A novel multistage system for the detection and removal of pectoral muscles in mammograms,” *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 26, no. 1, pp. 35–49, 2018.
- [25] S. A. Taghanaki, Y. Liu, B. Miles, and G. Hamarneh, “Geometry-based pectoral muscle segmentation from mlo mammogram views,” *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 11, pp. 2662–2671, 2017.
- [26] J. Wang and Y. Yang, “A context-sensitive deep learning approach for micro-calcification detection in mammograms,” *Pattern recognition*, vol. 78, pp. 12–22, 2018.
- [27] E. A. Rashed *et al.*, “Neural networks approach for mammography diagnosis using wavelets features,” *arXiv preprint arXiv:2003.03000*, 2020.
- [28] N. L. K. Sari, P. Prajitno, L. E. Lubis, and D. S. Soejoko, “Computer aided diagnosis (cad) for mammography with markov random field method with simulated annealing optimization,” *Journal of Medical Physics and Biophysics*, vol. 4, no. 1, pp. 85–94, 2017.

-
- [29] S. Sasikala, M. Ezhilarasi *et al.*, “Fusion of k-gabor features from medio-lateral-oblique and craniocaudal view mammograms for improved breast cancer diagnosis,” *Journal of cancer research and therapeutics*, vol. 14, no. 5, p. 1036, 2018.
- [30] C. Ancy and L. S. Nair, “Tumour classification in graph-cut segmented mammograms using glm features-fed svm,” in *Intelligent Engineering Informatics*. Springer, 2018, pp. 197–208.
- [31] L. Tsochatzidis, K. Zagoris, M. Savelonas, N. Papamarkos, I. Pratikakis, N. Arikidis, and L. Costaridou, “Microcalcification oriented content-based mammogram retrieval for breast cancer diagnosis,” in *2014 IEEE International Conference on Imaging Systems and Techniques (IST) Proceedings*. IEEE, 2014, pp. 257–262.
- [32] J. Wang, Y. Yang, M. N. Wernick, and R. M. Nishikawa, “An image-retrieval aided diagnosis system for clustered microcalcifications,” in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2016, pp. 1076–1079.
- [33] R. Nakayama, H. Abe, J. Shiraishi, and K. Doi, “Potential usefulness of similar images in the differential diagnosis of clustered microcalcifications on mammograms,” *Radiology*, vol. 253, no. 3, pp. 625–631, 2009.
- [34] L. Wei, Y. Yang, and R. M. Nishikawa, “Microcalcification classification assisted by content-based image retrieval for breast cancer diagnosis,” *Pattern recognition*, vol. 42, no. 6, pp. 1126–1132, 2009.
- [35] H. Jing, Y. Yang, and R. M. Nishikawa, “Retrieval boosted computer-aided diagnosis of clustered microcalcifications for breast cancer,” *Medical Physics*, vol. 39, no. 2, pp. 676–685, 2012.
- [36] W. Jian, X. Sun, and S. Luo, “Computer-aided diagnosis of breast microcalcifications based on dual-tree complex wavelet transform,” *Biomedical engineering online*, vol. 11, no. 1, pp. 1–12, 2012.
- [37] J. Nagi, S. A. Kareem, F. Nagi, and S. K. Ahmed, “Automated breast profile segmentation for roi detection using digital mammograms,” in *2010 IEEE EMBS conference on biomedical engineering and sciences (IECBES)*. IEEE, 2010, pp. 87–92.

-
- [38] M. Slavković-Ilić, A. Gavrovska, M. Milivojević, I. Reljin, and B. Reljin, “Breast region segmentation and pectoral muscle removal in mammograms,” *Telfor Journal*, vol. 8, no. 1, pp. 50–55, 2016.
- [39] C. Kanan and G. W. Cottrell, “Color-to-grayscale: does the method matter in image recognition?” *PloS one*, vol. 7, no. 1, p. e29740, 2012.
- [40] S. A. Agnes, J. Anitha, S. I. A. Pandian, and J. D. Peter, “Classification of mammogram images using multiscale all convolutional neural network (ma-cnn),” *Journal of medical systems*, vol. 44, no. 1, pp. 1–9, 2020.
- [41] K. A. H. Gómez, J. D. Echeverry-Correa, and Á. Á. O. Gutiérrez, “Automatic pectoral muscle removal and microcalcification localization in digital mammograms,” *Healthcare Informatics Research*, vol. 27, no. 3, p. 222, 2021.
- [42] N. Shrivastava and J. Bharti, “Breast tumor detection in digital mammogram based on efficient seed region growing segmentation,” *IETE Journal of Research*, pp. 1–13, 2020.
- [43] N. Alam and M. J. Islam, “Pectoral muscle elimination on mammogram using k-means clustering approach,” *International Journal of Computer Vision & Signal Processing*, vol. 4, no. 1, pp. 11–21, 2014.
- [44] M. Mustra and M. Grgic, “Robust automatic breast and pectoral muscle segmentation from scanned mammograms,” *Signal processing*, vol. 93, no. 10, pp. 2817–2827, 2013.
- [45] J. L. Rodríguez Cardona *et al.*, “Identificación de microcalcificaciones en mamografías usando redes neuronales convolucionales,” B.S. thesis, Uniandes.
- [46] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts, “Computational radiomics system to decode the radiographic phenotype,” *Cancer research*, vol. 77, no. 21, pp. e104–e107, 2017.
- [47] S. Haykin, *Neural networks and learning machines*, 3/E. Pearson Education India, 2010.
- [48] Y. Anzai, *Pattern recognition and machine learning*. Elsevier, 2012.
- [49] P. SUCKLING J, “The mammographic image analysis society digital mammogram database,” *Digital Mammo*, pp. 375–386, 1994.

-
- [50] C. Rose, D. Turi, A. Williams, K. Wolstencroft, and C. Taylor, “Web services for the dds and digital mammography research,” in *International workshop on digital mammography*. Springer, 2006, pp. 376–383.
- [51] A. Qayyum and A. Basit, “Automatic breast segmentation and cancer detection via svm in mammograms,” in *2016 International conference on emerging technologies (ICET)*. IEEE, 2016, pp. 1–6.
- [52] W. B. Yoon, J. E. Oh, E. Y. Chae, H. H. Kim, S. Y. Lee, and K. G. Kim, “Automatic detection of pectoral muscle region for computer-aided diagnosis using mias mammograms,” *BioMed research international*, vol. 2016, 2016.
- [53] K. Wei, W. Guangzhi, and D. Hui, “Segmentation of the breast region in mammograms using watershed transformation,” in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. IEEE, 2006, pp. 6500–6503.
- [54] S. Sreedevi and E. Sherly, “A novel approach for removal of pectoral muscles in digital mammogram,” *Procedia Computer Science*, vol. 46, pp. 1724–1731, 2015.
- [55] D. Raba, A. Oliver, J. Martí, M. Peracaula, and J. Espunya, “Breast segmentation with pectoral muscle suppression on digital mammograms,” in *Iberian conference on pattern recognition and image analysis*. Springer, 2005, pp. 471–478.
- [56] N. Saltanat, M. A. Hossain, and M. S. Alam, “An efficient pixel value based mapping scheme to delineate pectoral muscle from mammograms,” in *2010 IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA)*. IEEE, 2010, pp. 1510–1517.
- [57] L. Wang, M.-l. Zhu, L.-p. Deng, and X. Yuan, “Automatic pectoral muscle boundary detection in mammograms based on markov chain and active contour model,” *Journal of Zhejiang University SCIENCE C*, vol. 11, no. 2, pp. 111–118, 2010.
- [58] Y. Li, H. Chen, Y. Yang, and N. Yang, “Pectoral muscle segmentation in mammograms based on homogenous texture and intensity deviation,” *Pattern Recognition*, vol. 46, no. 3, pp. 681–691, 2013.
- [59] B. Graham, “Fractional max-pooling,” *arXiv preprint arXiv:1412.6071*, 2014.

-
- [60] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.
- [61] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [62] J. Wang, R. M. Nishikawa, and Y. Yang, “Global detection approach for clustered microcalcifications in mammograms using a deep learning network,” *Journal of Medical Imaging*, vol. 4, no. 2, p. 024501, 2017.
- [63] X. Liu, M. Mei, J. Liu, and W. Hu, “Microcalcification detection in full-field digital mammograms with pfc clustering and weighted svm-based method,” *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–13, 2015.
- [64] I. El-Naqa, Y. Yang, M. N. Wernick, N. P. Galatsanos, and R. M. Nishikawa, “A support vector machine approach for detection of microcalcifications,” *IEEE transactions on medical imaging*, vol. 21, no. 12, pp. 1552–1563, 2002.
- [65] R. Gallardo-Caballero, C. García-Orellana, A. García-Manso, H. González-Velasco, and M. Macías-Macías, “Independent component analysis to detect clustered microcalcification breast cancers,” *The Scientific World Journal*, vol. 2012, 2012.