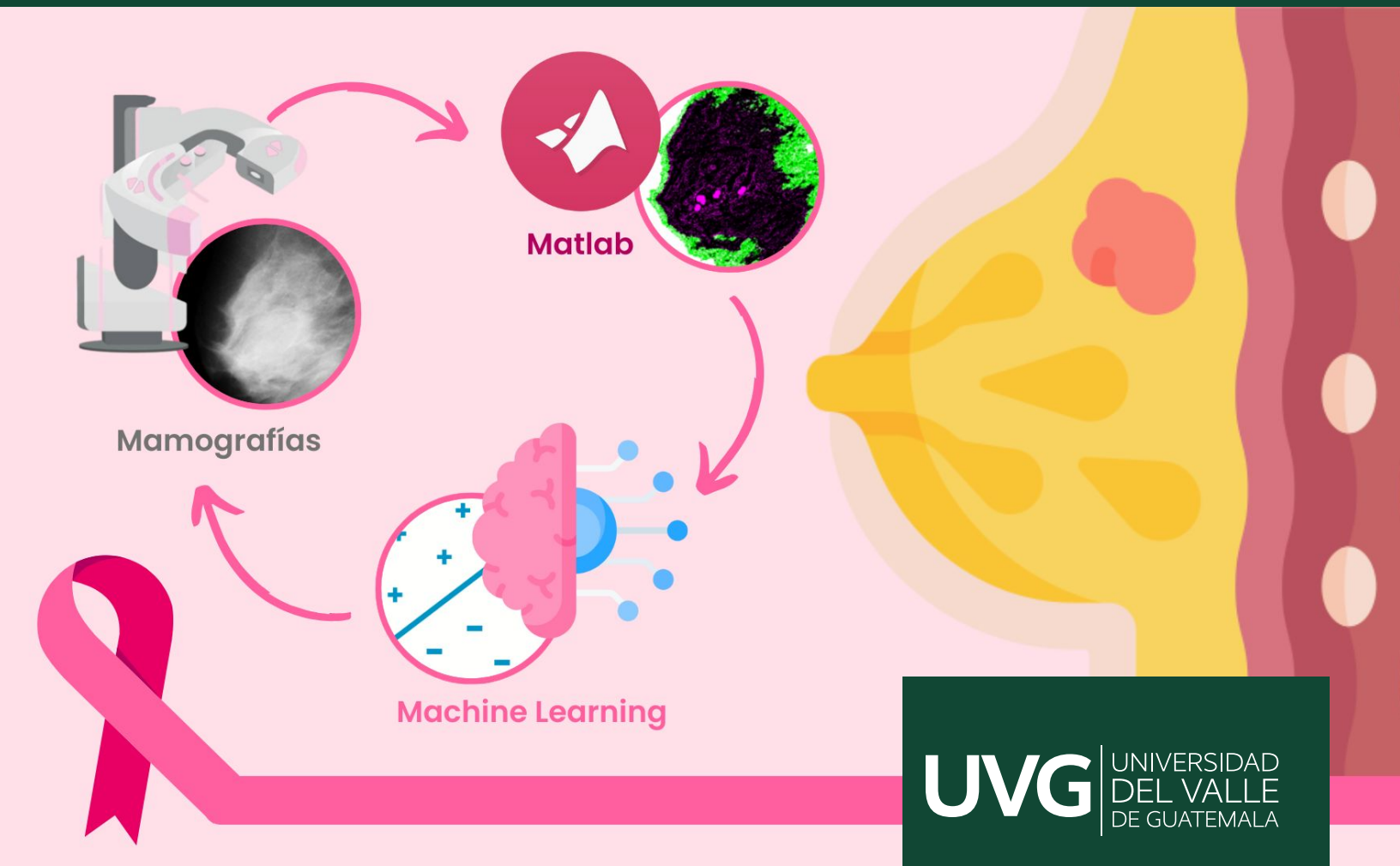
Categorización de tumores de mama a partir del procesamiento de imágenes médicas

Carmen Jimena Santizo Monterroso



UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería

Categorización de tumores de mama a partir del procesamiento de imágenes médicas

Trabajo de graduación presentado por Carmen Jimena Santizo Monterroso para optar al grado académico de Licenciado en Ingeniería Biomédica

Guatemala,

2024

UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería

Categorización de tumores de mama a partir del procesamiento de imágenes médicas

Trabajo de graduación presentado por Carmen Jimena Santizo Monterroso para optar al grado académico de Licenciado en Ingeniería Biomédica

Guatemala,

2024

Vo.Bo.:

(f) _____

M. Sc. Carlos Esquit

Tribunal Examinador:

(f) _____

M.Sc. Carlos Esquit

(f) _____

Dr. Luis Alberto Rivera Estrada

(f) _____

Ing. Kurt Emmanuel Kellner

Fecha de aprobación: Guatemala, 03 de diciembre de 2024.

Prefacio

El presente trabajo es el resultado de querer contribuir a mi país a través de la ingeniería biomédica. Entre las áreas de estudio, se destaca el de imagenología médica que en conjunto con *machine learning* conforman un área interdisciplinaria, enfocada en mejorar los sistemas de visualización y diagnóstico médico de manera no invasiva. Para el caso específico de mamografías, a pesar de que éstas suponen ser el estándar de oro para detectar el cáncer de seno, a través de una exploración diagnóstica por imagen, existe una gran complejidad en su interpretación. Esto debido a la variabilidad en la apariencia de las anormalidades, a las características del tejido mamario y a la calidad de las imágenes de mamografía. Por otra parte, la interpretación humana de una mamografía es subjetiva y cualitativa. Además, es evidente que en Guatemala el cáncer de mama se mantiene como la segunda causa de mortalidad para el país, con 2 mil 244 casos identificados para el año 2023 [13].

Por consiguiente, espero que esta tesis sea de gran utilidad como punto de partida para futuros investigadores interesados en el procesamiento de imágenes médicas, específicamente en estudios mamográficos para la detección temprana del cáncer de mama en países donde el sistema de salud es precario y requiere de atención.

Por ende, quiero expresar mi agradecimiento a quienes me han apoyado a lo largo de mi carrera universitaria. Primero, agradecer a Dios, por haber convertido todas mis batallas en retos llenos de aprendizajes, por haberme otorgado sabiduría y resiliencia y porque estoy segura que me continuará guiando. A mis padres y hermana, por su amor y apoyo incondicional, y por su excelente ejemplo de vida a seguir.

Agradezco a la Universidad del Valle de Guatemala por su compromiso con todos los estudiantes y por todas las oportunidades de crecimiento brindadas. Gracias a mi director de carrera M. Sc. Carlos Alberto Esquit Hernández y a mi asesor Dr. José Andrés Leal Ordóñez, quienes depositaron en mi la confianza y apoyo para hacer realidad este proyecto. Además, un especial agradecimiento al Hospital El Pilar, en particular al Dr. Aldo Mario Dardón Aguilar, quien mostró su interés y apoyo invaluable, contribuyendo éticamente en todo momento a través de su experiencia y profesionalismo.

Gracias a mis amigos, por confiar y creer en mí y haber hecho de mi etapa universitaria, un trayecto de vivencias que nunca olvidaré.

Índice

Lista de figuras

Lista de cuadros

Resumen

El cáncer de mama es una afección con gran prevalencia en el mundo, siendo la segunda causa de mortalidad por cáncer en mujeres. En Guatemala, es la neoplasia maligna más común; con pacientes jóvenes diagnosticadas en estadios avanzados y subtipos biológicamente agresivos. Por lo tanto, la detección precoz y temprana a través de imágenes médicas, supone la clave para incrementar las posibilidades de supervivencia. Sin embargo, el análisis y la interpretación de los estudios de imagenología médica se basa en una inspección visual con un valoración subjetiva y cualitativa, enfocada en evaluar la presencia de masas a través de una búsqueda de áreas con densidades y distribución del tejido anormales; lo que provoca posibles enmascaramientos y falsos positivos o negativos. Además, en el país los recursos son limitados y se requiere un cierto grado de atención especializada en los establecimientos de salud a nivel urbano como rural, así como a nivel público y privado; lo que implica un reto para el país.

En este proyecto se desarrolló con Matlab, una herramienta de apoyo para el diagnóstico basado en el aprendizaje automático para detectar calcificaciones mamarias y clasificarlas correctamente entre benignas y malignas. Tras aplicar un flujo de trabajo para mejorar la calidad de las mamografías, se hace uso de las características radiómicas del seno. Con ello se logra una precisión del 84 %, siendo un algoritmo capaz de hallar lesiones adicionales, además de una visualización de tumores como áreas concentradas en color magenta y en color verde las regiones sanas asociadas a los tejidos propios de la mama.

Abstract

Breast cancer is a highly prevalent condition in the world, being the second cause of cancer mortality in women. In Guatemala, it is the most common malignant neoplasm, with young patients diagnosed in advanced stages and biologically aggressive subtypes. Therefore, early detection through medical imaging is the key to increase the chances of survival. However, the analysis and interpretation of medical imaging studies is based on a visual inspection with a subjective and qualitative assessment, focused on evaluating the presence of masses through a search for areas with abnormal densities and tissue distribution; which causes possible masking and false positives or negatives. In addition, resources in the country are limited and a certain degree of specialized care is required in urban and rural health facilities, as well as public and private, which implies a challenge for the country.

In this project, a machine learning-based diagnostic support tool was developed with Matlab to detect breast calcifications and correctly classify them into benign and malignant. After applying a workflow to improve the quality of mammograms, the breast radiomic characteristics are used. With this an accuracy of 84 % is achieved, being an algorithm capable of finding additional lesions, in addition to a visualization of tumors as concentrated areas in magenta color and in green color the healthy regions associated with the breast tissues themselves.

CAPÍTULO 1

Introducción

Para el caso de países con un índice de desarrollo humano (IDH) bajo, como lo es Guatemala, si bien se diagnostica cáncer de mama a una de cada 27 mujeres, una de cada 48 muere [10]. Por ello, la exploración mamográfica se ha convertido en el pilar de los exámenes de detección y diagnóstico a nivel mundial [1] [10]. Por lo tanto, un análisis preciso supone una detección efectiva del cáncer de mama [10]. Pero si bien, el enfoque tradicional depende en gran medida de la experiencia de los médicos radiólogos y sus inspecciones visuales a partir de un aprendizaje natural basado en el reconocimiento de patrones [1], existen errores debido a las limitaciones y en especial cuando se deben asignar probabilidades a las observaciones.

En consecuencia, el uso de las herramientas de diagnóstico asistidas por computadora están destinadas a ayudar a los médicos a mejorar la precisión de los mismos [5]. Matlab es un lenguaje de programación que permite procesar y estudiar fácilmente las mamografías, facilitando la identificación de estructuras mamarias y la detección de anomalías [7]. Esto con el fin de auxiliar a los radiólogos a detectar lesiones [62] y proporcionar resultados clínicos rápidos y confiables que permitan complementar el diagnóstico médico. Además en conjunto con modelos de ML, es posible clasificar los tumores mamarios en benignos y malignos [5].

En los capítulos 2 y 3 se presentan los antecedentes más relevantes dentro del área de imagenología médica para la detección del cáncer de mama, seguido de un análisis de la incidencia de dicha neoplasia, y la justificación para aplicar el procesamiento de imágenes médicas en conjunto con *machine learning* para segmentar el cáncer de mama. Posteriormente, se explican los objetivos generales y específicos de este proyecto (capítulo 4) seguido de la delimitación de los mismos, tomando en cuenta los alcances y las limitaciones (capítulo 5). Luego, en el capítulo 6 se presenta el marco teórico que permite ahondar más acerca del cáncer de mama, diagnósticos y modelos de *machine learning*. En el capítulo 7 se muestra la metodología, detallando las fases del flujo de trabajo aplicado a cada una de las mamografías, para finalmente obtener una clasificación y visualización eficaz de lesiones mamarias. Los resultados se presentan en el capítulo 8, seguido de una discusión de estos (capítulo 9). Finalmente, en los capítulos 10 y 11 se establecen las conclusiones y recomendaciones respectivamente.

CAPÍTULO 2

Antecedentes

La imagenología médica mamaria permite efectuar diagnósticos de carcinoma de mama. Clínicamente, el diagnóstico se efectúa por medio de una mamografía, resonancia magnética o ecografía mamaria, se detecta y caracteriza las lesiones encontradas en el cribado (estrategia de prevención secundaria, para la detección precoz de una determinada enfermedad a fin de disminuir la incidencia de complicaciones derivadas de la patología). La mamografía es una imagen bidimensional y se basa en la identificación de hallazgos morfológicos sospechosos de cáncer de mama, incluyendo masas, asimetrías, calcificaciones agrupadas y áreas de distorsión arquitectónica [2].

La ecografía se utiliza especialmente en mujeres con mamas densas, ya que el cáncer puede estar oculto en la mamografía. También se emplea en pacientes con síntomas clínicos o anomalías palpables, pero con un examen mamográfico negativo. Por ende, el ultrasonido mamario permite analizar más a fondo una anomalía mamográfica, para determinar si una masa de tejido blando es sólida o quística y para diferenciar masas benignas de malignas [3].

La resonancia magnética (MRI) es una técnica valiosa para la obtención de imágenes mamarias, ya que permite evaluar la extensión de los tumores, detectar lesiones adicionales o cánceres que estaban ocultos en la mamografía o ecografía y establecer el tamaño del tumor, cuando éste difiere significativamente entre las modalidades de imagen [4].

No obstante, sin importar el método para la obtención de las imágenes médicas mamarias, el análisis y la interpretación del estudio es efectuado por un radiólogo a través de la inspección visual. Por lo tanto, la valoración del estudio es subjetiva, cualitativa y se basa en evaluar la presencia de nódulos, masas sólidas o quistes mamarios; a través de una búsqueda de áreas con densidades, tamaños, estructuras, formas, simetrías, distribución del tejido y bordes anormales [1].

Recientemente se demostró que la eficiencia y precisión del diagnóstico aumenta al momento de emplear algoritmos de aprendizaje automático. Mert et al. [5] señalan que un médico experimentado es capaz de realizar diagnósticos con aproximadamente 80 % de precisión, basándose únicamente en el análisis visual de las imágenes y sin tomar en cuenta

estudios o exámenes complementarios, como lo son las biopsias. Por el contrario, al emplear un sistema automático que hace uso de imágenes médicas, se logra un 91 % de diagnósticos correctos.

Es por ello que, en los últimos años la Inteligencia Artificial (IA) ha cobrado avances en la medicina a través de *machine learning* (ML), siendo posible clasificar lesiones mamarias en normales y anormales [6]. Asimismo, a través de la extracción de características de textura aplicados a pequeñas subregiones dentro del área de interés del ROI (parte de una imagen que se desea filtrar o procesar), se pueden obtener imágenes útiles para el médico en el diagnóstico y tratamiento del cáncer de mama [7]. Por otro lado, existe un aplicativo ejecutable en Matlab (Mathworks Inc, CA USA) con capacidad para obtener la estructura aproximada y una caracterización de la lesión de forma eficiente, a partir de imágenes ecográficas ingresadas por el operador del protocolo Breast Imaging-Reporting and Data System (BI-RADS) [8].

En consecuencia, se dice que los tumores benignos tienen forma redonda u ovalada, mientras que los tumores malignos tienen una forma parcialmente redondeada con un contorno irregular. Cabe destacar que la masa maligna aparecerá más blanca que cualquier tejido que la rodee [9]. Por consiguiente, las masas en el pecho pueden ser clasificadas como benignas o malignas dependiendo de su forma y características específicas aplicados a regiones de interés.

A partir de ello, se establece que la capacidad de automatizar la identificación y clasificación de un tumor, permite detectar tempranamente el cáncer de seno, conllevando a que los pacientes adquieran la terapia adecuada y se incrementen las posibilidades de supervivencia.

CAPÍTULO 3

Justificación

El cáncer de mama es una afección en la que las células cancerosas y con alteraciones de la mama, se multiplican sin control llegando a formar tumores que incluso son capaces de propagarse e invadir los ganglios linfáticos o los órganos cercanos al seno, causando metástasis e incluso la muerte [1]. De acuerdo a estadísticas de la Organización Mundial de la Salud (OMS), en 2020 aproximadamente 685 mil personas fallecieron por cáncer de mama en todo el mundo, siendo el 99 % de los casos en mujeres. Además, 2.3 millones de mujeres fueron diagnosticadas con cáncer de mama, lo que constituye a este tipo de cáncer como el de mayor prevalencia en el mundo y en la segunda causa de mortalidad por cáncer en mujeres [10].

Cabe destacar que aproximadamente el 50 % de los casos de cáncer de mama afectan a mujeres que no tienen factores de riesgo específicos aparte del sexo y la edad [10]. Sin embargo, actualmente la carga de enfermedad que representa el cáncer de mama es desproporcionadamente mayor en los países en vías de desarrollo, donde la mayoría de las muertes por cáncer de mama ocurren prematuramente, en mujeres menores de 70 años [11]. En el caso específico de Guatemala, la detección precoz y el acceso a tratamientos efectivos siguen siendo un reto, ya que los recursos médicos son limitados y se requiere un alto grado de atención especializada en muchos establecimientos de salud o centros de oncología tanto a nivel urbano como rural, así como a nivel público y privado. Asimismo, de acuerdo a estadísticas del año 2020, en Guatemala la mayoría de las mujeres con cáncer de mama residen en el departamento de Guatemala (46.3 %) y Escuintla (11.2 %), con una edad media en mujeres de 52.7 años y una prevalencia en estadios III (12.5 %) y II (7.5 %) [12].

Es evidente que en Guatemala el cáncer de pecho es la neoplasia maligna más común entre mujeres y es el segundo tipo de cáncer con más prevalencia y número de casos. Además de esto, las pacientes jóvenes con cáncer de mama en Guatemala se caracterizan por ser diagnosticadas en estadios avanzados de la enfermedad con subtipos biológicamente más agresivos como el triple negativo y el con fenotipo HER-2 enriquecido [13].

La carga de enfermedad por cáncer de mama se puede reducir mediante la identificación y el tratamiento temprano de los cánceres, antes de que den síntomas [11]. Asimismo, una

medida importante que se puede tomar para reducir el riesgo y evitar la muerte por cáncer de seno, consiste en la detección temprana. El cáncer de seno que se detecta precozmente, es más fácil de tratarlo y combatirlo, obteniendo buenos resultados y un índice bajo de recidivancia [10]. Sin embargo, muchos de los casos en los que se detecta cáncer de mama requieren de biopsias quirúrgicas, por estereotaxia, por aspiración o por punción [14]. Una biopsia mamaria a pesar de ser un procedimiento capaz de determinar si el tejido presenta células cancerosas, posee riesgos asociados como: hematoma, hinchazón, dolor, enrojecimiento o aumento de la temperatura de la mama, fiebre, infección, supuración inusual o sangrado en el sitio de la biopsia, alteración del aspecto y el proceso de cicatrización del seno y posibilidad de otra cirugía de acuerdo a los resultados del informe de patología [14]. Por ello, es preferible realizar procedimientos no invasivos que puedan proveer un alto grado de certeza sobre la presencia de un tumor maligno, antes de realizar un procedimiento invasivo.

Cabe destacar que las imágenes obtenidas a partir de ultrasonidos, rayos X o ecografías mamarias, deben de ser interpretadas y analizadas por radiólogos profesionales, lo cual representa un consumo de tiempo elevado y la posibilidad de error de interpretación humana, debido a que es una valoración subjetiva y cualitativa [1]. Típicamente, la clasificación de masas se basa en características de textura utilizando los contornos obtenidos mediante segmentación semiautomática [15]. Además de esto, las masas benignas tienden a tener límites redondos, mientras que las masas malignas suelen tener límites irregulares, por lo que el análisis de la forma permite una clasificación correcta de masas [9].

A partir del uso de las características radiómicas en conjunto con las herramientas de análisis y procesamiento de imágenes médicas, se eluden los procesos invasivos poco tolerantes para la obtención de una muestra de algunas células o tejidos, como lo son las biopsias; además de detectar tempranamente casos de patología mamaria, reducir el tiempo de análisis de resultados y mitigar el riesgo a error de interpretación por parte de los profesionales de salud. Por consiguiente, el uso de imágenes médicas en conjunto con *machine learning*, suponen una herramienta útil para obtener resultados rápidos y precisos en la clasificación, detección y segmentación del cáncer de mama.

Machine learning (ML) es ideal para la detección temprana de tumores, ya que las técnicas de clasificación inteligentes pueden ayudar a los médicos a identificar síntomas que tal vez no se observen mediante un enfoque convencional, debido a la naturaleza compleja de las microcalcificaciones y las masas [16]. Además, muchos investigadores han intentado aplicar algoritmos de aprendizaje automático para detectar la capacidad de supervivencia de los cánceres en seres humanos y también han demostrado que estos algoritmos funcionan mejor en la detección del diagnóstico de cáncer [17]. Por lo tanto, hoy en día el aprendizaje automático es ampliamente conocido como método de clasificación y modelado del carcinoma de mama. Es un método que puede encontrar regularidades y patrones ya oscuros a partir de una variedad de conjuntos de datos. Incorpora una amplia variedad de métodos utilizados para revelar reglas, paradigmas y conexiones en grupos de datos y produce una especulación de estas conexiones que pueden utilizarse para descifrar nuevos datos ocultos [18].

En este proyecto se busca detectar, clasificar y segmentar precoz y tempranamente el carcinoma de mama, por medio del procesamiento y entrenamiento de modelos de aprendizaje automático a partir de imágenes ecografías mamarias. De esta manera, se pretende reducir el número de muertes prematuras en mujeres afectadas por está enfermedad, además de disminuir costos y tiempo necesario para el tratamiento.

CAPÍTULO 4

Objetivos

4.1. Objetivo general

Crear un algoritmo de procesamiento de imágenes médicas mamográficas capaz de diferenciar entre tumores mamarios benignos y malignos.

4.2. Objetivos específicos

- Desarrollar algoritmo de preprocesamiento de imágenes médicas obtenidas de un repositorio, reduciendo los artefactos de la adquisición y estandarizando las imágenes para el análisis posterior.
- Identificar los mejores métodos para segmentar las regiones de interés en las imágenes preprocesadas a partir de las características radiómicas del seno, para la detección de masas mamarias.
- Entrenar algoritmos de *machine learning* con imágenes médicas obtenidas de repositorios para la detección y diferenciación automática entre tumores benignos y malignos.
- Obtener imágenes médicas de pacientes guatemaltecas y validar el funcionamiento de los algoritmos desarrollados en colaboración de expertos en radiología.

CAPÍTULO 5

Alcance

Si bien los objetivos de este proyecto giran en torno al procesamiento de imágenes médicas mamográficas para la clasificación de tumores de mama, no supone ser una herramienta médica de diagnóstico. Este algoritmo tiene como propósito ser una fuente de apoyo actuando como segunda opinión, donde inicialmente el médico radiólogo debe analizar y formular sus interpretaciones a partir de sus conocimientos y experiencia.

Cabe destacar que la calidad de las imágenes mamográficas es fundamental para la precisión, sensibilidad y especificidad del algoritmo. Por el contrario, las limitaciones del proyecto se establecen en base a la disponibilidad de un amplio banco de mamografías y metadatos asociados. La fiabilidad del algoritmo se evaluará por medio de regiones de interés (ROI) graficadas de acuerdo a las coordenadas y diámetros establecidos por el informe del *dataset*. Por lo tanto, será posible comparar y validar las lesiones encontradas experimentalmente, con las lesiones reales descritas. La eficacia se estimará con un conjunto de prueba que representa el 20 % de los datos analizados. La veracidad se analizará a partir de mamografías anonimizadas de pacientes mujeres guatemaltecas, empleando el modelo de *machine learning* creado a partir del conjunto de entrenamiento (80 % de los datos).

Por ende, este proyecto pretende aportar un sistema que ayude al médico radiólogo en la detección temprana y en el diagnóstico preciso, objetivo y eficaz de masas en mamografías digitales. Esto será posible a través de información de tipo radiómica que el algoritmo proporciona, indicando las posibles áreas en donde pueda existir una lesión.

Básicamente, el algoritmo se considera como un instrumento auxiliar en el diagnóstico y clasificación de lesiones mamarias, más no un sistema comercial para la detección de anormalidades mamarias. Además, se estima que sirva como un recurso de apoyo para categorizar y segmentar posibles enmascaramientos que no son perceptibles a través de una inspección visual y cualitativa. También se considera como una herramienta en los casos en los que las mamografías experimentan artefactos, ya sea por adquisición, compresión del seno o por la propia morfología, histología y anatomía de la mama.

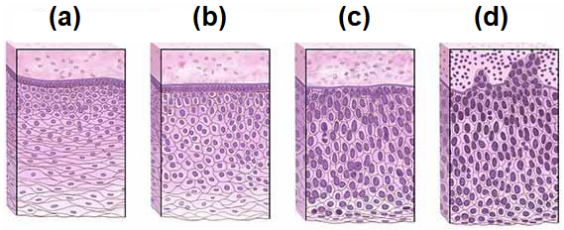
CAPÍTULO 6

Marco teórico

6.1. ¿Qué es el cáncer?

En condiciones normales, las células humanas se forman y se multiplican a través de la división celular a medida que el cuerpo las necesita. Sin embargo, a veces el proceso no sigue este orden y las células anormales o células dañadas se forman y se multiplican cuando no deberían, conllevando a la formación de tumores [19]. Existen cambios anormales (Figura 1) previo a que las células cancerosas se formen en los tejidos del cuerpo, como:

Figura 1. *a) Células normales, (b) hiperplasia, (c) displasia y (d) cáncer*



Modificado de [19].

- Hiperplasia: aumento anormal del número de células en un tejido del cuerpo, pero con apariencia normal.
- Displasia: también se acumulan demasiadas células, pero las células son anormales y la estructura del tejido tiene variaciones. No son cancerosas.
- Carcinoma *in situ*: enfermedad avanzada localizada. Presencia de células anormales que no se diseminan al tejido cercano. Se tratan, debido a que algunos carcinomas *in situ* se convierten en cáncer.

- Cáncer: multiplicación sin control de algunas células del cuerpo. Se considera cáncer metastásico, cuando el cáncer que se diseminó del sitio donde se inició a otra parte del cuerpo. Por lo que, el cáncer metastásico tiene el mismo nombre y el mismo tipo de células cancerosas que el cáncer primario [20].

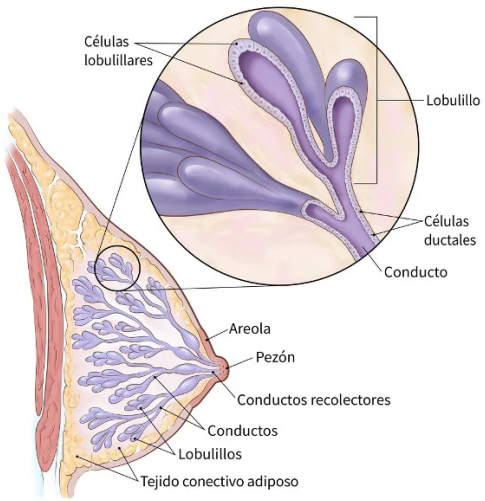
6.2. Carcinoma de mama

El seno es un órgano que se asienta sobre las costillas superiores y los músculos del pecho. El ser humano cuenta con dos senos, un seno izquierdo y uno derecho; siendo la cantidad de tejido graso lo que determina el tamaño de cada uno [21].

El seno femenino (Figura 2) está conformado por [21]:

- Lobulillos: glándulas encargadas de producir leche materna. Los cánceres que se originan en ésta zona se denominan cánceres lobulillares.
- Conductos: canales que salen de los lobulillos y son los encargados de transportar la leche materna hacia el pezón. Los cánceres que se generan en ésta parte son denominados cánceres ductales y son los más comunes.
- Pezón: abertura en la piel del seno, donde los conductos se unen y se convierten en conductos más grandes capaces de permitir la salida de la leche del seno. Se encuentra rodeado de una piel ligeramente más oscura y gruesa, llamada areola.
- Estroma: conformado por tejido adiposo y conectivo, vasos sanguíneos, vasos linfáticos y nervios. Aporta nutrientes al tejido y elimina el exceso de residuos y líquido.

Figura 2. *Anatomía de la mama femenina*



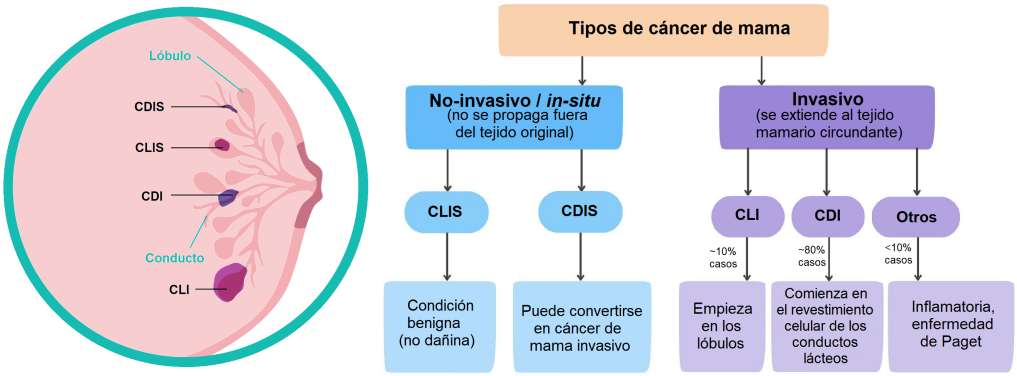
Obtenido de [22].

Por lo que, el cáncer mamario (también referido como cáncer de mama o de seno) es un proceso oncológico en el que células sanas de la glándula mamaria, se degeneran y se transforman en tumorales [21]. Es una enfermedad, que se da a través de una proliferación acelerada e incontrolada de células del epitelio glandular [23] [24] y que en uno de sus tipos prolifera hasta constituir un tumor, que posteriormente invade tejidos circundantes y hace metástasis a distintas áreas del cuerpo a partir de los vasos sanguíneos o los ganglios linfáticos [25].

6.2.1. Tipos de cáncer de mama

Existen diversos tipos de cáncer de seno (Figura 3), pero es la clase específica de células afectadas en el seno [21] [24], lo que permite determinar el tipo de cáncer [26]. Sin embargo, comúnmente se clasifican entre no invasivo/*in situ* e invasivo:

Figura 3. *Tipos de cáncer de mama no invasivos e invasivos*



Nota. La ubicación del tumor es importante para el diagnóstico, ya que el tipo de cáncer de mama depende de en qué parte de la mama se formó el tumor.

El cáncer de mama no invasivo o *in situ*, es el estadio más precoz del cáncer de mama. Afecta una zona importante del seno, pero no ha invadido los tejidos circundantes y no se ha diseminado hacia otras partes del cuerpo [24]. Un ejemplo de este tipo de cáncer incluye al cáncer lobulillar *in situ* (CLIS), afección mamaria inocua y benigna que se desarrolla en los lobulillos, pero no se considera carcinoma. Otro tipo de cáncer no invasivo es el ductal *in situ* (CDIS), también conocido como carcinoma intraductal o cáncer de seno en etapa 0. Es un tipo de cáncer no invasivo o preinvasivo, en donde las células que revisten los conductos son ahora células cancerosas, pero no se han propagado por las paredes de los conductos hasta el tejido mamario adyacente [21]. Este tipo representa el 85 % de los carcinomas in situ y al menos la mitad de los cánceres de mama. En la mayoría de los casos, se detecta mediante mamografía [24].

Por el contrario se encuentra el cáncer de mama invasivo, también conocido como infiltrante debido a que las células cancerosas se han propagado al tejido mamario cercano. Un ejemplo de este tipo de cáncer es el carcinoma lobulillar invasivo o infiltrante (CLI), el cual comienza en las glándulas mamarias que producen leche (lobulillos) [24]. Al igual

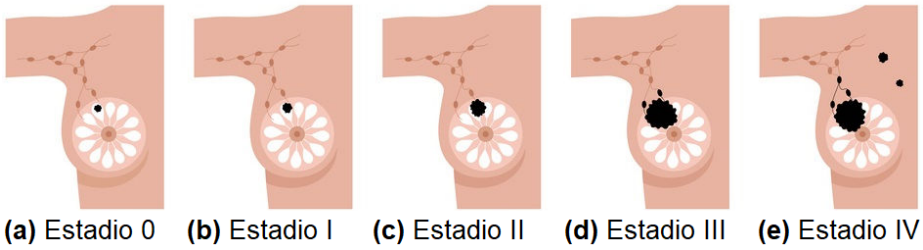
que el CDI, es capaz de provocar metástasis a otras partes del cuerpo. Es más probable que afecte ambos senos, con una incidencia en 1 de cada 5 mujeres [21] [26]. Sin embargo, también se encuentra el carcinoma ductal invasivo (CDI), el tipo de cáncer de seno más común, representando alrededor de un 75 % de los cánceres de mama invasivos [24]. Comienza en las células que revisten el conducto de leche, hasta la pared del conducto y los tejidos mamarios cercanos. Tiene la capacidad de provocar metástasis a través del sistema linfático y el torrente sanguíneo [21]. Además, está el cáncer de seno inflamatorio (IBC), el cual es un tipo de CDI y es causado por células cancerosas que bloquean los vasos linfáticos en la piel, lo que provoca inflamación, hinchazón y enrojecimiento en el seno. No parece un cáncer de seno típico, ya que a menudo no causa un tumor mamario y por ende, es más difícil de diagnosticarlo [21]. Finalmente, está la enfermedad de Paget, una formación anormal de células en la piel del pezón y la areola [21].

También existen otros tipos de cánceres de mama pero poco comunes: medular, mucinoso, tubular, metaplásico, papilar y triple negativo (TNBC) [24].

6.2.2. Estadificación del cáncer de mama

La estadificación o determinación de la etapa del cáncer se determina en función de las características y la presencia o no de receptores hormonales [24]. Para el caso específico del cáncer de seno, éste suele expresarse con un número de estadio entre 0 y IV (Figura 4). El estadio 0 corresponde a los tipos de cáncer no invasivos o *in situ* que permanecen en su ubicación original, mientras que el estadio IV a los tipos de cáncer invasivos que hacen metástasis fuera de la mama [24] [27].

Figura 4. *Estadios del cáncer de mama. (a) Estadio 0: tumor no invasivo menor de 2 cm, (b) estadio I: tumor invasivo mayor a 2 cm, (c) estadio II: tumor que mide de 2 a 5 cm, (d) estadio III: tumor que mide más de 5 cm y (e) estadio IV: tumor que se disemina a otros órganos, tejidos o partes del cuerpo*



Modificado de [27].

El 'sistema TNM' de la American Joint Committee on Cancer (AJCC) es el procedimiento de estadificación del cáncer de mama [27], que permite describir los casos de cáncer de manera uniforme. El estadio se calcula en función de parámetros T, N y M [23].

La categoría T, indica el tamaño o extensión del tumor primario, permitiendo determinar si hay metástasis (invasión) en tejido cercano o no [23]. Los números de T más altos significan un tumor más grande y una propagación más extensa a los tejidos cerca del seno. Por ende,

una nomenclatura TX representa que no se puede evaluar el tumor primario, T0 no hay indicio alguno de tumor primario, Tis el tumor se encuentra in situ, hay presencia de CDIS o existe enfermedad de Paget del seno pero sin masas tumorales asociadas. T1 (T1a, T1b y T1c) el tumor es de 2 cm (3/4 de pulgada) o menos de ancho. T2 el tumor de más de 2 cm, pero no más de 5 cm (2 pulgadas) de ancho. T3 el tumor de más de 5 cm de ancho. T4 (T4a, T4b, T4c, y T4d) el tumor de cualquier tamaño y crece hacia la pared torácica o la piel. Incluye al cáncer de seno inflamatorio [27].

El parámetro N indica la presencia o no del tumor en ganglios (nódulos) linfáticos adyacentes [23]. Por lo tanto, NX representa que los ganglios linfáticos adyacentes no se pueden evaluar. N0 implica que el cáncer no se ha propagado a los ganglios linfáticos adyacentes, N1 el cáncer se propagó hacia 1 y 3 ganglios linfáticos axilares (debajo del brazo) o internos y N2 el cáncer se ha propagado a 4 y 9 ganglios linfáticos debajo del brazo, o el cáncer ha agrandado los ganglios linfáticos mamarios internos. La categoría N3 cuenta con dos posibles escenarios: N3a que implica que el cáncer se ha propagado a 10 o más ganglios linfáticos axilares con por lo menos un área de propagación del cáncer que mide más de 2 mm; o el cáncer se ha propagado a los ganglios linfáticos infraclaviculares (debajo de la clavícula) con por lo menos un área de propagación del cáncer que mide más de 2 mm. Mientras que N3b significa que hay existencia de cáncer en por lo menos 1 ganglio linfático axilar (con por lo menos un área de propagación del cáncer que mide más de 2 mm), y ha agrandado los ganglios linfáticos mamarios internos; o el cáncer se ha propagado a 4 o más ganglios linfáticos axilares (con por lo menos un área de propagación del cáncer que mide más de 2 mm), y hacia los ganglios linfáticos mamarios internos en la biopsia de ganglio linfático centinela [27].

En cuanto al parámetro M, éste indica si el cáncer se ha propagado o no a los órganos distantes (metástasis) [23]. M0 hace referencia a que no se encuentra propagación a distancia en estudios por imágenes o por examen médico, mientras que M1 implica que el cáncer se ha propagado a órganos distantes y mide más de 0.2 mm [27].

Cabe mencionar que en 2018, la AJCC actualizó los estándares de estadificación de cáncer de mama, incluyendo etapas clínicas y patológicas [27]. A partir de ello, se agregaron otras características [23] a tomar en cuenta para determinar el estadío de una manera más precisa, las cuales son:

- Estado del receptor de estrógeno (ER).
- Estado del receptor de progesterona (PR).
- Estatus proteína HER2.
- Grado del cáncer (G): ¿qué tanto las células cancerosas se parecen a las células normales?

Una vez que se han determinado todos los factores del sistema de estadificación, la información se combina en un proceso llamado agrupación por etapas para asignar una etapa general [23] y poder estadificar la lesión.

6.2.3. Lesiones mamarias

A pesar de que existen los estándares de estadificación de cáncer de mama, el Colegio Estadounidense de Radiología (ACR) estableció un sistema para uniformar las descripciones que usan los radiólogos en los resultados de una mamografía [28].

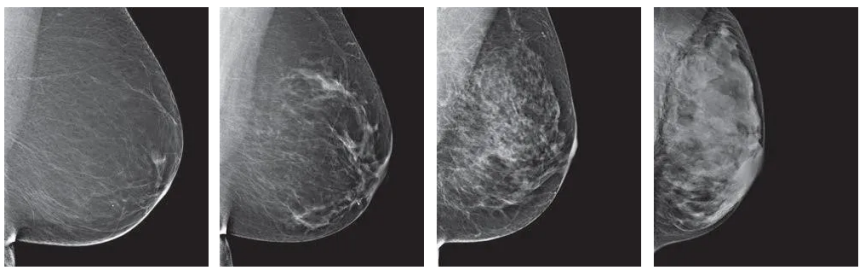
El sistema se conoce como BI-RADS y consta de 7 categorías estándar:

- Categoría 0: evaluación incompleta, se necesitan pruebas de imagen adicionales antes de asignar una categoría.
- Categoría 1: evaluación normal, presencia de lesiones mamarias negativa.
- Categoría 2: lesión benigna, no cancerosa.
- Categoría 3: lesión probablemente benigna.
- Categoría 4: lesión sospechosa, probablemente maligna.
- Categoría 5: altamente indicativo de cáncer.
- Categoría 6: cáncer confirmado mediante biopsia.

Además, el sistema BI-RADS incluye cuatro categorías para clasificar la densidad de la mama [28]. Las categorías de mayor a menor densidad (Figura 5) de la mama son:

- Las mamas cuentan con una composición casi por completo de grasa.
- Hay áreas dispersas de tejido fibroglandular (glandular y conjuntivo).
- Las mamas son de densidad heterogénea, con mayor áreas de densidad fibroglandular.
- Las mamas son de densidad extrema. Esto quizás oculte tumores en el tejido y hace que sea más difícil verlos en la mamografía.

Figura 5. *Densidad de los senos a partir de imágenes obtenidas por medio de mamografía: (a) seno casi con puro tejido adiposo (grasa), (b) seno con algunas áreas de tejido denso glandular y tejido fibroso, (c) la mayor parte del seno se conforma de tejido denso glandular y tejido fibroso (descrito como heterogéneamente denso) y (d) seno extremadamente denso*

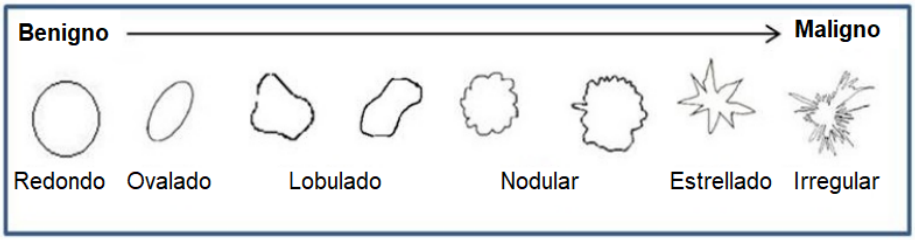


Modificado de [22].

Por consiguiente, los trastornos de la mama pueden ser benignos (no cancerosos) o malignos (cancerosos). La mayoría no son cancerosos y a menudo, no necesitan tratamiento [24]. Las lesiones benignas son totalmente asintomáticas, generalmente no necesitan ser extirpados, no invaden los tejidos circundantes y no se diseminan a otras partes del cuerpo. Se caracterizan por consistencia blanda, límites definidos, dolor frecuente, piel sin cambios y con movilidad. Por el contrario, las lesiones malignas son nódulo independiente al tejido mamario que lo circunda, o engrosamiento de una zona de la mama sin bordes claros. Son capaces de invadir otras partes del cuerpo humano y pueden aparecer nuevamente luego de su extirpación. Se caracterizan por consistencia dura, límites irregulares, movilidad fija, dolor poco frecuente y piel con edemas o retracciones.

Por ende, una diferencia clave entre los tumores benignos y malignos es su forma (Figura 6). Los tumores irregulares representan tumores malignos, mientras que los tumores regulares son benignos.

Figura 6. *Diferencias entre tumores benignos y malignos. Los tumores con formas más irregulares son malignos, por otro lado, los tumores con límites redondos, regulares y lisos se encuentran en etapa benigna*



Modificado de [29].

6.2.4. Epidemiología

La incidencia de cáncer de mama en América Latina varía ampliamente de un país a otro en función de su estatus socioeconómico, cobertura de atención médica, disponibilidad de procedimientos de detección estándar y métodos de diagnóstico [30]. Sin embargo, la incidencia de este tumor maligno está aumentando en todas las regiones del mundo debido al estilo de vida y la genética [31].

Además cabe destacar que a pesar de que el cáncer de seno ocurre casi exclusivamente en las mujeres, también puede ocurrir en hombres pero con una incidencia de solamente el 1 % [21] [23]. En las mujeres el cáncer de mama es el segundo tipo de cáncer más común a nivel mundial, seguido por el cáncer de piel y pulmón [19] [24]; así como la primera causa de muerte por tumores malignos [32]. Se estima que 2,089 millones de mujeres fueron diagnosticadas con cáncer de mama en 2018 y que actualmente, las pacientes con cáncer de mama representan hasta el 36 % de las pacientes oncológicas [33]. Aproximadamente un nuevo caso de cáncer de seno es diagnosticado cada 18 segundos [26].

En 2023, de acuerdo a [24] en Estados Unidos se estimaron 297,790 nuevos casos de cáncer femenino de mama invasivo, 55,720 nuevos casos de cáncer de mama no invasivo (*in*

situ) y 43,700 muertes por cáncer de mama; mientras que en hombres, 2,800 nuevos casos de cáncer de mama invasivo y 530 muertes a causa de este.

6.2.5. Factores de riesgo

La causa inequívoca de la carcinogénesis mamaria aún no se ha establecido, pero se conocen varios factores de riesgo que conducen a su desarrollo [32]. Entre los factores de riesgo, se encuentran aquellos que no son modificables y modificables.

Los factores no modificables incluyen [25] [31] [34]:

- Sexo: el sexo femenino constituye uno de los principales factores asociados con un mayor riesgo de cáncer de mama, principalmente debido a la mayor estimulación hormonal de estrógeno (ER) [34]. La gran mayoría de los casos de cáncer de mama que alcanzan el 99 %, se producen en mujeres [32].
- Edad: en la actualidad, alrededor del 80 % de los pacientes con cáncer de mama son personas de más de 50 años, mientras que al mismo tiempo más del 40 % son mayores de 65 años [23] [32] [34].
- Antecedentes familiares: aproximadamente el 13-19 % de las pacientes con cáncer de mama cuentan con un familiar de primer grado (hermana, madre o hija) afectado por la misma afección [34]. El riesgo aumenta significativamente con un número cada vez mayor de familiares de primer grado afectados [23] [33].
- Mutaciones genéticas: solo un pequeño grupo de casos de cáncer de mama (5-10 %) son genéticos, asociados con mutaciones en los genes BRCA1 (gen ubicado en el cromosoma 17) y BRCA2 (gen ubicado en el cromosoma 13) [24] [33].
- Raza/etnia: las mujeres blancas no hispanas cuentan con una tasa de incidencia de cáncer más alta, pero con una tasa de mortalidad mayor en mujeres con ascendencia africana [24]. Sin embargo, son éstas últimas las que se caracterizan por tener una tasa de supervivencia más baja [34].
- Historial reproductivo: estudios indican que el riesgo de desarrollar eventos cancerígenos en el microambiente mamario aumenta en proporción al estado hormonal de una mujer, la exposición al estrógeno y progesterona y la ocurrencia de eventos reproductivos específicos como el embarazo y la lactancia [23] [32] [35].
- Densidad del tejido mamario: en general, a mayor densidad de tejido mamario, mayor riesgo de cáncer de mama [23] [34].
- Antecedentes de cáncer de mama y enfermedades benignas de la mama: los antecedentes personales de cáncer de mama se asocian con un mayor riesgo de lesiones cancerosas y no cancerosas dentro de las mamas [23] [24].
- Radioterapia previa: el riesgo de neoplasias malignas secundarias tras el tratamiento radioterápico está estrictamente asociado con la edad del individuo [34]; ya que las pacientes que reciben radioterapia antes de los 30 años tienen un mayor riesgo de padecer cáncer de mama [23] [32] [36].

Por otro lado, los factores modificables se encuentran relacionados con el medio ambiente o estilos de vida, tales como [25] [31]:

- Terapia hormonal: las mujeres que usan terapia de reemplazo hormonal (TRH), especialmente durante más de 5 o 7 años, también tienen un mayor riesgo del 24 % de padecer cáncer de mama [23] [34].
- Actividad física: se considera que la actividad física regular es un factor protector de la incidencia de cáncer de mama [34].
- Índice de masa corporal (IMC): según la evidencia epidemiológica, la obesidad se asocia con una mayor probabilidad de cáncer de mama [34].
- Ingesta de alcohol: el consumo de alcohol contribuye al aumento de la concentración de estrógenos (ER) en la sangre al inhibir el metabolismo en el hígado e intensificar la conversión de andrógenos en estrógenos [34].
- Tabaquismo: los carcinógenos presentes en el tabaco son capaces de transportarse al tejido mamario, lo que aumenta la plausibilidad de mutaciones dentro de los oncogenes y los genes supresores [33] [34].
- Alimentación: se ha demostrado que existe una relación significativa entre el padecimiento de cáncer de mama y el consumo excesivo de carne roja y procesada, grasas saturadas y sodio [34]. Sin embargo, se sabe que la relación entre el cáncer de mama y la dieta es compleja, multifactorial y no lineal [32].
- Otras drogas: el uso de anticonceptivos orales aumenta muy levemente el riesgo (5 casos por cada 100,000 mujeres) de padecer cáncer de seno durante los años de uso. Sin embargo, el riesgo disminuye durante los 10 años siguientes a su interrupción [23].

Además, debido a que la incidencia del cáncer de mama aumenta constantemente en todas las regiones del mundo, parece necesario buscar nuevos métodos terapéuticos, así como factores predictivos y de pronósticos que se traduzcan en una disminución de las tasas de mortalidad [32].

6.2.6. Terapia y tratamientos

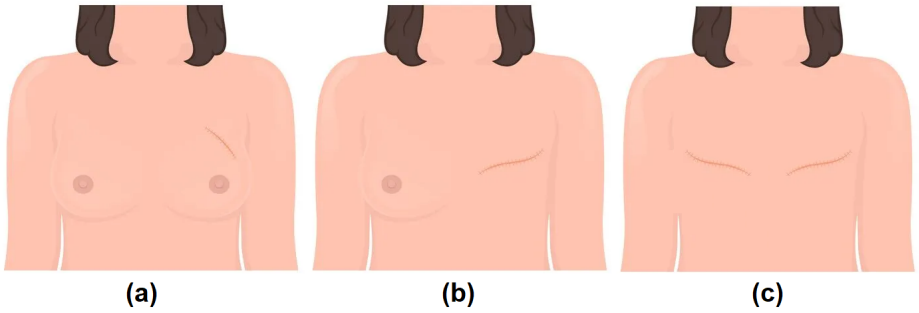
Las estrategias de tratamiento del cáncer de mama son multidisciplinarias, ya que dependen de: la etapa y extensión del cáncer, características histopatológicas, la velocidad de diseminación, categoría del estadio clínico [23], perfil de biomarcadores, estado de salud del paciente, expresiones de factores de crecimiento, entre otros [32]; ya que diferentes poblaciones de células madre y células progenitoras en la glándula mamaria pueden causar un cambio de paradigma en la comprensión de la heterogeneidad [26].

Sin embargo, entre los procedimientos quirúrgicos y terapias (Figura 7) comúnmente utilizados para tratar el carcinoma de mama [23] [32] se encuentran:

Figura 7. *Tratamientos y terapias para tratar el cáncer de seno*

- Radioterapia: destrucción de células cancerosas a partir de rayos X de alta energía u otros tipos de radiación. La forma en que se administra la radioterapia depende del tipo y el estadio del cáncer que se esté tratando [37].
- Quimioterapia: combinación de medicamentos fuertes capaces de destruir todas las células cancerosas, evitar que el cáncer se disemine hacia los ganglios linfáticos y disminuir el riesgo de que el cáncer aparezca nuevamente [37].
- Terapia hormonal: uso de medicamentos que bloquean las hormonas estrógeno (ER) y progesterona (PR); para reducir o destruir las células cancerosas que son sensibles a los tipos de cáncer de mama. Controla la diseminación hacia otras partes del cuerpo y reduce el riesgo de que el cáncer reaparezca [37].
- Inmunoterapia: tratamiento con medicamentos que ayudan al sistema inmunológico del paciente a destruir las células cancerosas específicas [37].
- Cirugía: procedimiento para la extirpación del cáncer de mama o de algunos ganglios linfáticos cercanos. Entre las cirugías (Figura 8) se encuentran [23] [32]:

Figura 8. *(a) Mastectomía simple, (b) mastectomía total y (c) mastectomía profiláctica contralateral*



- Mastectomía simple: extirpación solamente de una parte del tejido mamario, conservando otras partes del seno, como el pezón o la piel.
- Mastectomía total: extirpación del tejido mamario, lóbulos, conductos, pezón, piel y areola; es decir, eliminación por completo de la mama.
- Tumorectomía / lumpectomía / nodulectomía: extirpación del cáncer y parte del tejido sano que lo rodea. El resto del tejido mamario se conserva.
- Extirpación del ganglio linfático centinela: permite analizar si el cáncer se diseminó a los ganglios linfáticos cercanos; a partir de la extirpación de algunos

ganglios. Si no se encuentra cáncer en ninguno de éstos, no es necesario extirpar ningún otro.

- Extirpación del sistema linfático de la axila: solamente se da cuando los estudios por imágenes médicas indican que el cáncer se diseminó hacia los ganglios linfáticos o cuando se descubre la presencia de cáncer a partir de una biopsia de ganglios linfáticos centinela. Por lo general, se extirpan de 10 a 20 ganglios linfáticos axilares [24].
- Extirpación de ambos senos o mastectomía profiláctica contralateral: se da en casos de alto riesgo a desarrollar cáncer en la otra mama.

Por lo tanto, la American Cancer Society (ACS) sostiene que generalmente es probable que la cantidad de tratamiento necesario (local o sistémico) es proporcional a la extensión y etapa misma del cáncer de seno [23].

6.3. Diagnóstico del cáncer de mama

El cáncer de mama es una afección que aqueja en especial a muchas mujeres alrededor del mundo. Por ende, el uso de equipo médico que proporciona imágenes médicas es esencial para su diagnóstico. La resonancia magnética (RM) y la ecografía han demostrado su capacidad para detectar cáncer de mama, pero la sensibilidad y especificidad de la mamografía frente a ambos equipos representa mayor eficacia como estándar de oro en mujeres sin síntomas. De acuerdo a un metaanálisis, se demuestra que la mamografía continúa siendo la técnica más efectiva en mujeres sin riesgo elevado, mientras que la RM es empleada como un procedimiento de imagenología médica complementaria en mujeres con alto riesgo. Además, la ecografía muestra una tasa de detección más baja en comparación con la RM [38].

En consecuencia, se establece que la principal forma de detectar el cáncer de seno en estadios tempranos es a través del uso de un mamograma que permite obtener imágenes mamográficas (mamografías). A pesar de que la RM es útil en mujeres con riesgo elevado de cáncer, no se recomienda como reemplazo de la mamografía para el cribado rutinario en la población general debido a una tasa elevada de falsos positivos. Además, aunque existen tecnologías emergentes como la tomosíntesis, la mamografía sigue siendo la herramienta más confiable y accesible para el cribado rutinario de cáncer de seno, categorizándose como la modalidad más efectiva en la reducción de la mortalidad por cáncer de mama, especialmente en el grupo de mujeres de mediana edad (50 a 69 años) [39].

6.3.1. Mamografía

La mamografía de cribado es el principal método más utilizado en todo el mundo para la detección precoz de neoplasia mamaria en mujeres asintomáticas, y es la única modalidad de imagen que ha logrado reducir significativamente la mortalidad [40]. Se ha demostrado que las mujeres que se someten a exámenes de detección con mamografía tienen una reducción de aproximadamente el 30 % en la mortalidad, en comparación con las mujeres que no se someten a pruebas de detección [41]. Sin embargo, sólo entre el 10 y el 15 % de las

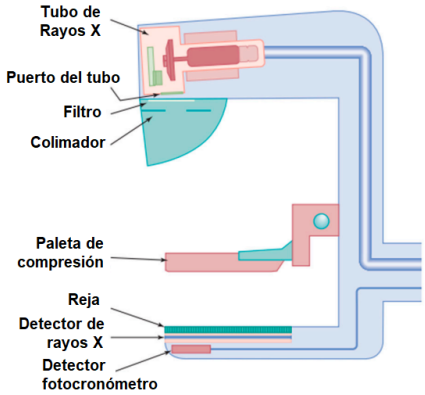
alteraciones detectadas durante una mamografía de cribado rutinaria resultan ser un cáncer [24]. Por ende, de acuerdo a la mayoría de las directrices internacionales de cribado, las mamografías son valiosas para las mujeres de 40 años en adelante [23]. En el caso específico de la ACS, ésta se recomienda que las mujeres de 45 a 54 años de edad se realicen pruebas de detección anuales, mientras que las mujeres de 55 o más años realicen una transición a las pruebas de detección bienales. Lo antes mencionado es debido a que, la tasa de crecimiento del cáncer de seno es más rápida en las mujeres premenopáusicas. Por el contrario, en las mujeres posmenopáusicas, aunque el beneficio máximo se logra con los exámenes de detección anuales, el beneficio incremental de ese enfoque en comparación con los exámenes de detección bienales es menos marcado [40].

El mayor valor de la mamografía se observa en el grupo de mujeres de 50 a 69 años, ya que se caracteriza por una sensibilidad del 75-95 % y una especificidad del 80-95 % [33]. Éste último parámetro establece que la mamografía permite declarar negativa a una paciente que está realmente libre de cáncer de mama con una alta fiabilidad [42]. Para las mujeres con sospecha de cáncer de mama hereditario, la mamografía por resonancia magnética se utiliza como prueba de detección [32]. No obstante, la tasa de falsos positivos es significativamente alta, ya que un aumento en la densidad del tejido mamario fibroglandular aumenta la posibilidad de enmascarar o imitar una lesión subyacente en una mamografía; porque tanto el tejido denso como el cáncer parecen blancos [43]. Además, otro problema al que se enfrenta el cribado es el sobrediagnóstico: el diagnóstico de neoplasias de mama indolentes que no se habrían hecho clínicamente evidentes durante la vida del paciente [41].

6.3.2. Principios físicos de la mamografía

Aún no se ha diseñado la unidad de mamografía perfecta desde el punto de vista del mamógrafo ni de la mujer. Pero en términos generales, el equipo básico de mamografía (Figura 9) consta de: consola de control, monitor, generador, tubo de rayos X, filtro de rayos X, colimador, placas detectoras, dispositivo de exposición automática y paletas de compresión [42].

Figura 9. *Sistema y partes de mamograma básico*

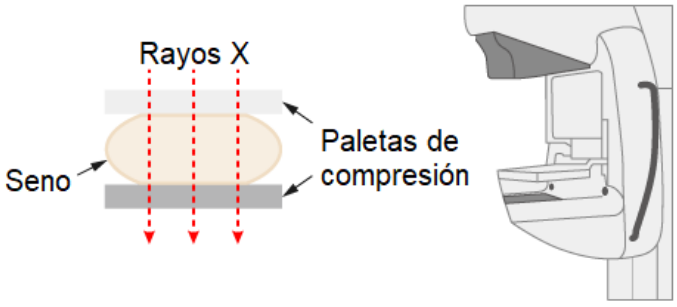


Modificado de [42].

En comparación con los equipos de imagen de rayos X, un mamógrafo está equipado con una placa de compresión motorizada, generalmente de policarbonato, que puede ser de hasta 100 N. La compresión es necesaria para reducir la cantidad de dosis administrada y obtener imágenes de alta calidad: al reducir el grosor por el que pasa la radiación, también disminuye el número de fotones dispersos que llegan al detector [42]. Por lo tanto, los principios físicos que rigen una mamografía, son los mismos que los de la radiografía (rayos X) pero en una forma especializada [44].

Los rayos X son un tipo de radiación electromagnética ionizante que permite obtener una imagen radiopaca cuando los rayos X logran impactar en menor medida en la placa, mientras que se obtiene una imagen radiolúcida cuando la estructura interpuesta deja pasar más rayos X (Figura 10). Por lo tanto, en una mamografía se establecen parámetros de imagen específicos para optimizar el contraste de los tejidos de la mama [44], ya que existen pequeñas diferencias entre los tejidos mamarios normales y aquellos en los que hay masas tumorales: la atenuación de los rayos X en el tejido glandular es similar a la del tejido adiposo [42]. Entonces, en una mamografía los rayos X examinan los bultos en los senos, los cambios en la piel y la secreción o engrosamiento del pezón [39]; siendo capaz de revelar masas/lesiones, calcificaciones, tumores y distorsiones arquitectónicas en el tejido mamario. Sin embargo, a pesar de ser el estándar de oro en el cribado de mamas, tiene limitaciones significativas [43].

Figura 10. *Irradiación de la mama con rayos X. La mama se estira entre dos placas delgadas y los tejidos de alta densidad absorben fácilmente los rayos X, mientras que en los tejidos de baja densidad los rayos X penetran*



Modificado de [45].

6.3.3. Principios de absorción de energía durante una mamografía

El funcionamiento de un mamograma se basa en el principio de absorción de energía a partir de la radiación. Considerando dos tipos de tejido en la mama: graso y blando, un mamograma crea contraste entre estos dos tejidos porque atenúan (es decir, dispersan y absorben) diferentes cantidades de radiación. Como resultado, diferentes cantidades de radiación llegan al detector debajo de cada tejido, lo que hace que se muestren diferentes niveles de gris en la mamografía (se crea contraste en la imagen) [42]. Sin embargo, es importante tomar en cuenta que la cantidad de radiación que llega al detector debajo de cada trozo de tejido depende de varios factores relacionados con la composición del tejido

y el haz de rayos X. Los tejidos con un número atómico más alto (es decir, que contienen elementos con más protones) tienen una atenuación sustancialmente mayor debido a una mayor cantidad de interacciones fotoeléctricas [44].

El calcio en las microcalcificaciones tiene un número atómico de 20, que es más del doble que el carbono, el hidrógeno y el oxígeno que componen los tejidos blandos y la grasa. Este alto número atómico da lugar a más eventos fotoeléctricos, aumentando sustancialmente la atenuación de los rayos X [44]. El tejido con mayor densidad física también bloquea una mayor cantidad de radiación. El aumento de la densidad física significa que también hay una mayor densidad de electrones por centímetro cúbico en el tejido. Dado que los rayos X interactúan con los electrones, el aumento de la densidad de electrones implica un mayor número de interacciones y, por lo tanto, una mayor atenuación [42]. Del mismo modo, la probabilidad de que un rayo X interactúe (se atenúe) aumenta cuanto más tejido atraviesa. En consecuencia, una lesión gruesa será más atenuante que una lesión delgada con la misma composición [44].

Finalmente, la atenuación de todo el tejido disminuye a medida que aumenta la energía del haz de rayos X. Por ende, es más importante que a medida que aumenta la energía de los rayos X, la diferencia de atenuación entre los tejidos disminuya. Esto significa que a medida que aumenta la energía de los rayos X, el número de fotones que llegan al detector debajo de dos tejidos diferentes se vuelve similar, lo que lleva a valores de nivel de gris similares en la imagen (bajo contraste de imagen). En consecuencia, en la mamografía se deben utilizar energías de haz muy bajas para generar un contraste de imagen suficiente para la detección de lesiones. Sin embargo, si la energía del haz es demasiado baja, los fotones no penetrarán en el pecho para golpear el detector. Esto da como resultado una imagen ruidosa y una dosis alta de radiación en la mama [44].

6.3.4. Tipos de mamografías

El cribado del cáncer con mamografía se considera eficaz para reducir la mortalidad relacionada con el cáncer de mama [44]. Existen 3 tipos de mamografías. El primer tipo es la mamografía mejorada con contraste (CEM), siendo una técnica de diagnóstico por imágenes que utiliza medio de contraste yodado para mejorar la visualización de las lesiones mamarias y la evaluación de la neovascularización tumoral [46]. Debido a que es una técnica con modificaciones en la energía de los rayos X, se obtienen imágenes con resaltes en áreas de absorción y acumulación de medios de contraste. La CEM tiene la ventaja de demostrar tanto cambios anatómicos como cambios locales en la perfusión mamaria, presumiblemente causados por la angiogénesis tumoral [44].

El segundo tipo de mamografía es la digital de campo completo (FFDM). Es una radiografía de la mama que utiliza un detector digital acoplado a una computadora digital, además de técnicas de procesamiento de imágenes digitales en 2 dimensiones (2D) para mejorar la visibilidad de los detalles y el contraste de la imagen [42]. Por el contrario, se encuentra la tomosíntesis digital de mama (DBT), que permite adquirir múltiples proyecciones a través de una trayectoria predefinida para obtener imágenes tridimensionales (3D) y seccionales de la mama [47]. Estas imágenes seccionales intentan superar la limitación de la superposición de tejidos, especialmente en mamas densas donde es difícil detectar algunas

lesiones mamarias [41]. Sin embargo, a pesar de los avances tecnológicos y los distintos tipos de mamografías, la visibilidad (detección) de las lesiones en mamas muy densas sigue siendo controvertida para el cribado por mamografías [48].

6.3.5. Otros estudios complementarios

Existen otros estudios por imágenes médicas [39] que permiten observar el tejido mamario y con ello, detectar la presencia o no de anomalías, lesiones o tumores. Tal es el caso de la ecografía mamaria, un método que utilizar ondas ultrasónicas para capturar a través de imágenes, las estructuras internas de las mamas e incluso, determinar si es un tumor sólido o un quiste (líquido). Las ondas rebotan en los tejidos u órganos y producen ecos. Los ecos forman la imagen del tejido que se llama ecograma [37]. También existe la resonancia magnética (RM), con un funcionamiento basado en el campo magnético y ondas de radio para generar imágenes más detalladas del interior de las mamas [37] [41].

Sin embargo, en algunos casos es necesario tomar una muestra de tejido de la mama (biopsia) [24] [39] para analizarla y posteriormente, confirmar si hay cáncer o no. La biopsia es un procedimiento que consiste en introducir una aguja a través de la piel, hasta llegar al tejido mamario y extraer una muestra del mismo para su análisis [24]. Existen diferentes tipos de biopsias mamarias: por escisión, por incisión, por punción con aguja gruesa o por aspiración con aguja fina (AAF) [37].

6.4. *Machine learning (ML)*

El aprendizaje automático o ML es un campo de estudio de la inteligencia artificial (IA) que se basa en enfoques estadísticos para dotar a los ordenadores de la capacidad de 'aprender' de los datos [49] con el fin de mejorar su rendimiento para resolver tareas sin estar programados explícitamente para cada una de ellas y hace predicciones o decisiones basadas en datos pasados [4] [50] [51].

Machine learning (ML) emplea dos tipos de técnicas: aprendizaje supervisado y no supervisado. El aprendizaje supervisado entrena un modelo con datos de entrada y salida conocidos para predecir salidas futuras, mientras que el aprendizaje no supervisado identifica estructuras intrínsecas o patrones ocultos en los datos de entrada y se emplea para sacar conclusiones sobre conjuntos de datos de entrada sin respuestas etiquetadas [52].

Para el caso del aprendizaje supervisado, éste cuenta con dos técnicas: de clasificación y regresión. Las técnicas de clasificación clasifican los datos de entrada en categorías y predicen respuestas discretas, como por ejemplo, si un tumor es benigno o maligno. Las técnicas de regresión predicen respuestas continuas. Por el contrario, la agrupación en *clusters* es la técnica más común del aprendizaje no supervisado, empleada para el análisis exploratorio de datos, a fin de identificar patrones o grupos ocultos en los datos. Los algoritmos de agrupación en *clusters* se dividen en dos grupos principales: agrupación en *clusters* dura, donde cada punto de datos pertenece a un solo *cluster* y agrupación en *clusters* blanda, donde cada punto de datos puede pertenecer a más de un *cluster* [52].

6.4.1. ¿Cómo se aplica ML a imágenes médicas mamarias?

En el sistema de atención de la salud, ha habido un aumento dramático en la demanda de servicios de imágenes médicas. Sin embargo, éstas a menudo son difíciles de analizar y es un proceso lento debido a la escasez de radiólogos y a los artefactos presentes en las mismas [4] [51]. Sin embargo, seleccionar el algoritmo adecuado específicamente para imágenes médicas puede resultar abrumador, ya que existen una infinidad de modelos y cada uno adopta un enfoque de aprendizaje distinto. Por ende, encontrar el algoritmo adecuado es cuestión de ensayo y error, pero siempre se debe de tomar en cuenta el tamaño y tipo de datos con que se trabaje, la información que se desee obtener de los datos y cómo se empleará [52].

Por lo tanto, se dice que el proceso de ML no suele ser y rara vez es lineal de principio a fin, ya que éste consiste en realizar iteraciones constantemente y probar distintos enfoques e ideas. Sin embargo, cada flujo de trabajo de ML comienza con 3 preguntas: ¿con qué tipo de datos va a trabajar?, ¿qué información desea obtener de ellos? y ¿cómo y dónde se aplicará esa información?. Las respuestas a estas preguntas ayudarán a decidir si conviene utilizar aprendizaje supervisado o no supervisado. Se selecciona el aprendizaje supervisado si se necesita entrenar un modelo para realizar una predicción y si se dispone de datos existentes para la salida que se desea predecir. Mientras que se selecciona el aprendizaje no supervisado si se necesita explorar datos y desea entrenar un modelo para obtener una representación interna [52].

En consecuencia, el algoritmo de aprendizaje supervisado toma un conjunto conocido de datos de entrada (el conjunto de entrenamiento) y respuestas conocidas sobre esos datos (salidas), y entrena un modelo para generar predicciones razonables como respuesta a datos de entrada nuevos. Todas las técnicas de aprendizaje supervisado son básicamente una forma de clasificación o regresión. Las técnicas de clasificación predicen respuestas discretas y se entrenan para clasificar datos en categorías. Entre sus aplicaciones se incluyen captura de imágenes médicas. Las técnicas de regresión predicen respuestas continuas [52].

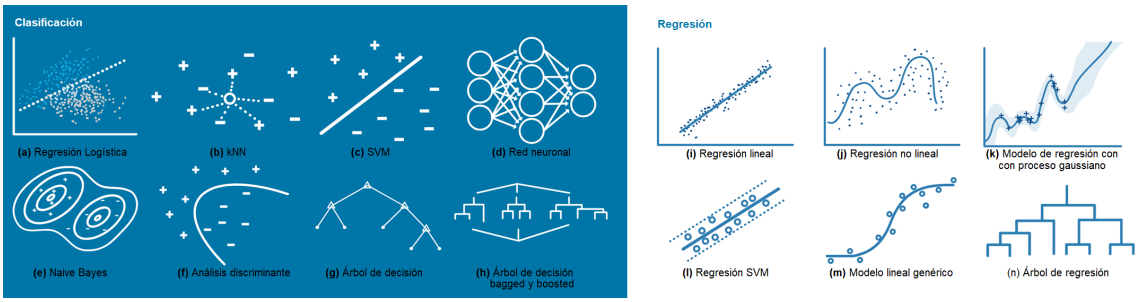
El algoritmo de aprendizaje no supervisado resulta útil si desea explorar los datos pero aún no tiene un objetivo específico o no sabe a ciencia cierta qué información contienen los datos. También es un buen modo de reducir las dimensiones de los datos. La mayoría de las técnicas de aprendizaje no supervisado son básicamente una forma de análisis de *clusters*. En el análisis de *clusters*, los datos se dividen en grupos según un grado de confianza o pertenencia. Los *clusters* se forman de modo que los objetos en un mismo *cluster* tienen características muy similares, y los objetos en diferentes *clusters* tienen características totalmente distintas [52]. No obstante, las principales técnicas o métodos que se utilizan para el análisis del conjunto de datos (diagnósticos) de cáncer de mama son: de regresión, máquinas de vectores de soporte, árboles de decisión y redes neuronales [49].

6.4.2. Modelos

Entre los principales modelos de ML para las técnicas de aprendizaje supervisado [49], tanto para clasificación y regresión (Figura 11), aplicados a imágenes médicas están máquina de vectores de soporte (SVM), análisis discriminante, Naive Bayes, K vecino más cercano, regresión lineal, GLM, SVR, GPR, métodos de ensemble, árboles de decisión y redes neuro-

nales.

Figura 11. *Algoritmos de clasificación y regresión para aprendizaje supervisado*



Modificado de [52].

La máquina de vectores de soporte (SVM) clasifica los datos buscando el límite de decisión lineal (hiperplano) que separa todos los puntos de datos de una clase de los de la otra clase. El mejor hiperplano para una SVM es aquel que tiene el margen más amplio entre las dos clases, cuando los datos se pueden separar linealmente. Si los datos no se pueden separar linealmente, se utiliza una función de pérdida para penalizar puntos situados en el lado incorrecto del hiperplano. A veces, las SVM emplean una transformación de kernel para convertir datos no linealmente separables en dimensiones superiores donde se puede encontrar un límite de decisión lineal. Se recomienda el uso de SVM para datos que tengan exactamente dos clases (también se puede utilizar para clasificación multiclase con una técnica denominada códigos de salida de corrección de error). También se recomienda emplearlo para datos de alta dimensionalidad, no linealmente separables o cuando se necesita un clasificador sencillo, fácil de interpretar y preciso [52].

El análisis discriminante clasifica los datos buscando combinaciones lineales de características, por lo que presupone que diferentes clases generan datos basados en distribuciones gaussianas. Entrenar un modelo de análisis discriminante implica hallar los parámetros de una distribución gaussiana para cada clase. Los parámetros de distribución se utilizan para calcular límites, que pueden ser funciones lineales o cuadráticas. Estos límites se emplean para determinar la clase de los datos nuevos. Se recomienda su uso cuando se necesita un modelo sencillo fácil de interpretar, cuando el uso de memoria durante el entrenamiento es motivo de preocupación o cuando se necesita un modelo que realice predicciones rápidamente. Por otro lado, un clasificador Naive Bayes presupone que la presencia de una determinada característica en una clase no está relacionada con la presencia de otra característica. Clasifica datos nuevos en base a la mayor probabilidad de que pertenezcan a una determinada clase. Se recomienda su uso para un conjunto de datos pequeño que contenga muchos parámetros, cuando se necesita un clasificador fácil de interpretar o cuando el modelo se encontrará con escenarios que no forman parte de los datos de entrenamiento, como es el caso de muchas aplicaciones médicas [52].

K vecinos más cercanos (kNN) categoriza los objetos en función de las clases de sus vecinos más cercanos en el conjunto de datos. Las predicciones presuponen que los objetos cercanos entre sí, son similares. Para hallar el vecino más cercano, se utilizan las métricas

de distancia euclidiana, distancia Manhattan, distancia del coseno y distancia de Chebychev. Se recomienda su utilización cuando se necesita un algoritmo sencillo para establecer reglas de aprendizaje de referencia, cuando el uso de memoria del modelo entrenado es de menor preocupación o cuando la rapidez de predicción del modelo entrenado es de menor preocupación. La regresión logística ajusta un modelo que puede predecir la probabilidad de que una respuesta binaria pertenezca a una clase u otra. Debido a su simplicidad, se utiliza comúnmente como punto de partida para problemas de clasificación binaria. Se recomienda su uso cuando los datos se pueden separar claramente con un solo límite lineal o como base de referencia para evaluar métodos de clasificación más complejos [52].

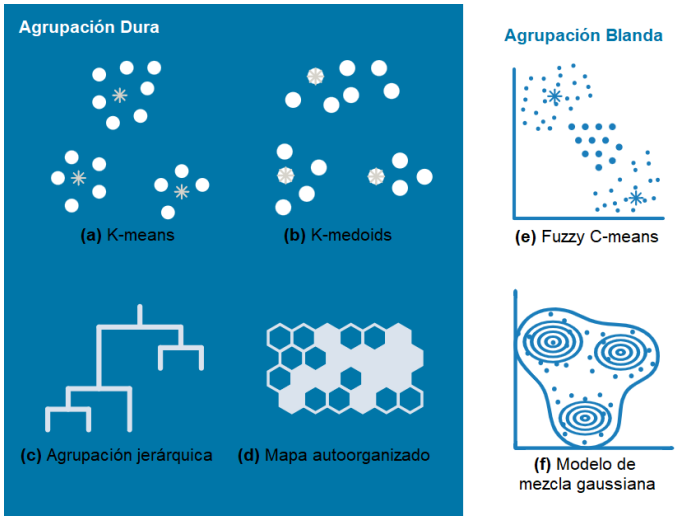
Entre los métodos de ensemble se encuentran los árboles de decisiones *bagged* y *boosted*, que combinan varios árboles de decisión 'débiles' en un conjunto más 'fuerte'. Un árbol de decisión *bagged* consta de árboles que se entrenan de manera independiente con datos que se extraen de los datos de entrada mediante *bootstrapping*. Por su parte, el método *boosted* implica crear un *learner* fuerte agregando de manera iterativa *learners* 'débiles' y ajustando el peso de cada *learner* débil para centrarse en ejemplos mal clasificados. Se recomienda aplicarlo cuando los predictores son categóricos (discretos) o se comportan de manera no lineal, o cuando el tiempo necesario para entrenar un modelo es de menor preocupación. Por otra parte, un árbol de decisión permite predecir las respuestas a datos siguiendo las decisiones en el árbol desde la raíz (inicio) hasta un nodo hoja. Un árbol consta de condiciones de ramificación en las que el valor de una variable de predicción se compara con un peso entrenado. El número de ramas y los valores de los pesos se determinan en el proceso de entrenamiento. Se puede realizar una modificación adicional, o poda, para simplificar el modelo. Se recomienda su uso cuando se necesita un algoritmo fácil de interpretar y rápido de ajustar, para reducir el uso de memoria o cuando no se requiere una alta precisión predictiva. Además, también se encuentra la red neuronal.Está inspirada en el cerebro humano, por lo que consta de redes de neuronas altamente conectadas que relacionan las entradas con las salidas deseadas. La red se entrena modificando de manera iterativa los pesos de las conexiones para que las entradas proporcionadas se correspondan con la respuesta correcta. Se recomienda utilizar red neuronal para mmodelar sistemas altamente no lineales, cuando los datos están disponibles gradualmente y se desea actualizar el modelo constantemente, cuando podría haber cambios inesperados en los datos de entrada o cuando la interpretabilidad del modelo no es una preocupación principal [52].

Entre los algoritmos de regresión comunes se encuentra la regresión lineal. La regresión lineal es una técnica de modelado estadístico que se emplea para describir una variable de respuesta continua como función lineal de una o varias variables de predicción. Su uso se recomienda cuando se necesita un algoritmo fácil de interpretar y rápido de ajustar o como base de referencia para evaluar otros modelos de regresión más complejos. También se encuentra la regresión no lineal, la cual también es una técnica de modelado estadístico y ayuda a describir relaciones no lineales entre datos experimentales. Los modelos de este tipo se suelen considerar paramétricos, es decir, que se describen como ecuaciones no lineales. Se recomienda su uso cuando los datos tienen tendencias no lineales fuertes y no se pueden transformar fácilmente en un espacio lineal, o para ajustar modelos personalizados a datos. Por otra parte está el modelo de regresión SVM que funciona como los algoritmos de clasificación SVM, pero se modifican para predecir una respuesta continua. En lugar de buscar un hiperplano que separe los datos, éste busca un modelo que se desvíe de los datos medidos por un valor no superior a una pequeña cantidad, con valores de parámetros lo más

pequeños posible (para reducir la sensibilidad al error). Se recomienda utilizarlo para datos de alta dimensionalidad, donde existe un gran número de variables de predicción [52].

En cuanto a los algoritmos de ML para las técnicas de aprendizaje no supervisado (Figura 12) de agrupación dura [49], se encuentran: *K-means*, *K-medoids*, agrupación jerárquica y mapa autoorganizado. Para la agrupación blanda, se menciona: *Fuzzy C-means* y modelo de mezcla gaussiana.

Figura 12. *Algoritmos de agrupación dura y blanda para aprendizaje no supervisado*



Modificado de [52].

K-means divide los datos en un número k de *clusters* mutuamente excluyentes. La distancia desde ese punto hasta el centro del *cluster* determina cúan bien encaja un punto en un *cluster*. Se recomienda utilizarlo cuando se conoce el número de *clusters* o para agrupar rápidamente grandes conjuntos de datos en *clusters*. *K-medoids* es similar a *K-means*, pero con el requisito de que los centros de *clusters* deben coincidir con puntos de datos. Se recomienda aplicarlo cuando se conoce el número de *clusters*, para agrupar rápidamente datos categóricos en *clusters* o para escalar a grandes conjuntos de datos. La agrupación jerárquica produce conjuntos anidados de *clusters* mediante el análisis de similitudes entre pares de puntos y la clasificación de objetos en un árbol jerárquico binario. Se recomienda emplearlo cuando no se sabe de antemano cuántos *clusters* hay en los datos o cuando se desea que la visualización guíe la selección. Y el mapa autoorganizado es una agrupación en *cluster* basada en red neuronal que transforma un conjunto de datos en un mapa en 2D que conserva la topología. Se recomienda su uso para visualizar datos de alta dimensionalidad en 2D o 3D, o para deducir la dimensionalidad de datos conservando su topología (formato) [52].

Para el caso de los algoritmos de agrupación blanda comunes, se destaca el *Fuzzy C-means*; que es una agrupación en *cluster* basada en particiones cuando los puntos de datos pueden pertenecer a más de un *cluster*. Se recomienda su uso cuando se conoce el número de *clusters*, para reconocimiento de patrones o cuando los *clusters* se superponen. Y el modelo de mezcla gaussiana, que es una agrupación en *cluster* basada en particiones donde los

puntos de datos provienen de diferentes distribuciones normales multivariantes con ciertas probabilidades. Se recomienda aplicarlo cuando un punto de datos podría pertenecer a más de un *cluster* o cuando los *clusters* tienen diferentes tamaños y estructuras de correlación internas [52].

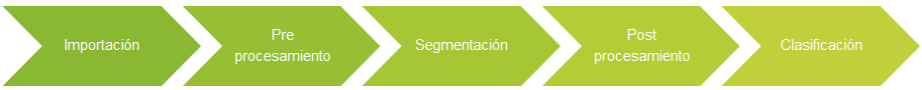
CAPÍTULO 7

Metodología

Para el análisis de las imágenes mamográficas médicas, es necesario aplicar un flujo de trabajo en Matlab (Mathworks Inc., CA, USA). La secuencia de procesos se lleva a cabo con imágenes obtenidas a partir de una base de datos preexistente en internet: Mammographic Image Analysis Society (MIAS, UK) database v1.21 [53]. El banco de datos consiste en 322 imágenes originales, siendo 161 pares de películas en donde la mamografía con un número de referencia impar representa la imagen de la mama izquierda y la de número par, de la mama derecha de una misma paciente. El *dataset* cuenta con una resolución de 50 micrones en formato en escala de grises PGM (*portable graymap*) y datos de veracidad asociados por medio de un informe. Dichas imágenes corresponden a mamografías digitales con tomas bilaterales, medio laterales y oblicuas; además de la presencia de masas normales o con anomalías tanto cancerígenas como no cancerígenas.

El flujo de trabajo inicia con la importación de las imágenes médicas, para luego efectuar un preprocesamiento, segmentación, posprocesamiento y una clasificación correcta de masas en dos clases: B - Benigno y M - Maligno (Figura 13).

Figura 13. *Flujo de trabajo para imágenes médicas*



Cabe destacar que también se forma una alianza con el Departamento de Radiología del Hospital El Pilar, con el propósito de obtener imágenes mamográficas reales y anonimizadas que permitan evaluar la precisión del modelo al momento de clasificar en las dos clases de tumores existentes.

7.1. Importación

La primera fase se basa en la importación de las imágenes médicas digitales con el fin de visualizarlas, cargarlas y almacenarlas en el espacio de trabajo de Matlab y extraer información de los diferentes tipos y datos de imágenes. Sin embargo, debido a que el *dataset* corresponde a un total de 322 mamografías, se automatiza la importación de las mismas a través de una búsqueda y un filtro que itera sobre cada archivo en formato .pgm existente en el directorio actual del *workspace* de Matlab.

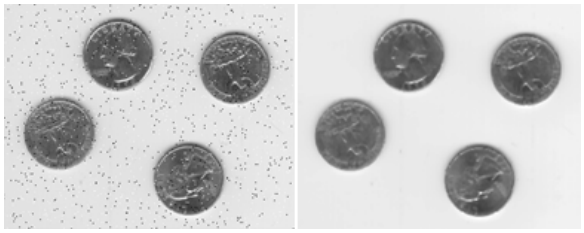
Además, debido a que la base de datos cuenta con datos asociados en un archivo con formato *.pdf*, se recurre a una condensación de la información para cada mamografía en un archivo de tipo *.xlsx*. A partir del archivo para hojas de cálculo, con cada iteración que realiza el programa automáticamente sobre cada mamografía, se extrae la información necesaria.

7.2. Preprocesamiento

La etapa de preprocesamiento implica efectuar ajustes a las imágenes médicas, reduciendo cualquier artefacto de adquisición. Este paso incluye la eliminación de fondo y ruido, normalización de la intensidad y ajuste del contraste. Inicialmente se convierten las imágenes mamográficas a escala de grises (*rgb2gray*), con el objetivo de que toda imagen esté en el formato adecuado para el análisis o la manipulación posterior con solamente un canal, a través del cual se representa la intensidad de la luz en diferentes niveles de gris.

En particular, se aplica un filtro de mediana (*medfilt2*), el cual se utiliza a menudo como parte del preprocesamiento y filtrado de artefactos en imágenes médicas, reduciendo el ruido 'sal y pimienta'. El tratamiento detallado con anterioridad permite la obtención de imágenes médicas homogeneizadas, estandarizadas, de mayor calidad, con reducción de distorsión y preservación de agudeza visual; para su posterior análisis en aplicaciones médicas de diagnóstico (Figura 14).

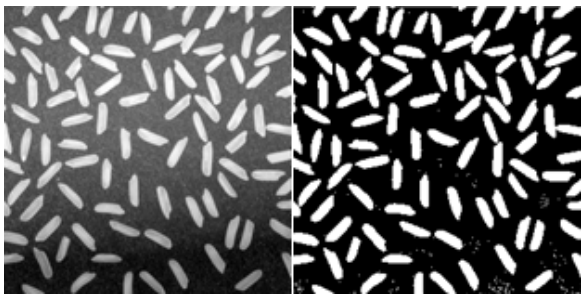
Figura 14. *Filtrado de mediana en 2D*



Obtenido de [54].

Se efectúa la conversión de las imágenes médicas en escala de grises a binarias (*imbinarize*), minimizando la varianza de los píxeles en base a un valor límite estimado. Por ende, la binarización de una imagen provoca que ésta se almacene como una matriz lógica con únicamente dos valores posibles para cada píxel, siendo 0 para píxeles negros y 1 para píxeles blancos (Figura 15).

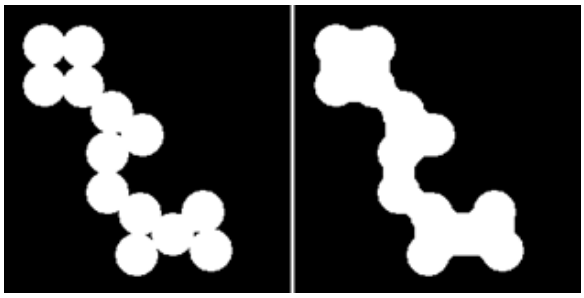
Figura 15. *Binarización de imagen en escala de grises*



Obtenido de [55].

Otra operación morfológica aplicada es el aislamiento del fondo (cierre de la imagen), a través de la cual se obtienen unos bordes más definidos y circunscritos y se resaltan determinados patrones o formas de la imagen binaria por medio de operaciones morfológicas (Figura 16). El cierre morfológico (*imclose*) de la imagen consiste en una dilatación seguida de una erosión. La dilatación se utiliza para expandir o engrosar las regiones de las imágenes al agregar píxeles a los bordes del área de interés médico, mientras que la erosión se aplica para reducir o adelgazar las regiones al eliminar píxeles en los bordes de la región deseada. Para ambas operaciones, se utiliza un mismo elemento estructurante morfológico (*strel*) que asume valores binarios.

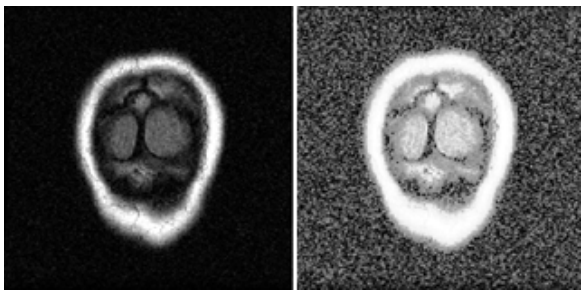
Figura 16. *Cierre morfológico para relleno de huecos en imagen*



Obtenido de [56].

Se mejora el contraste global de la imagen médica, a partir de la ecualización del histograma (*histeq*). En consecuencia, *histeq* permite obtener una imagen con valores de píxeles distribuidos uniformemente en todo el intervalo de intensidad, maximizando el contraste pero sin perder información de tipo estructural y con un mismo número de píxeles para cada nivel de gris del histograma de una imagen monocroma (Figura 17).

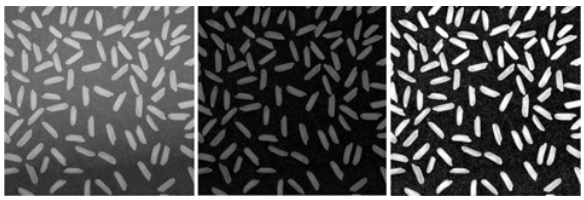
Figura 17. *Mejora del contraste mediante ecualización de histograma*



Obtenido de [57].

Con base en lo anterior, se aplica un filtro morfológico *top-hat*. El filtrado *top-hat* se aplica a la imagen binaria con el comando *imtophat*, permitiendo una corrección de la iluminación irregular y un resalte de detalles claros en la presencia de sombras o fondos oscuros (Figura 18). Al igual que con el filtro *imclose*, éste utiliza un elemento estructurante morfológico en forma de disco; pero la forma geométrica del entorno varía según el propósito del análisis de la imagen médica. Luego, a partir del comando *imadjust* se ajustan los valores de intensidad de la imagen tomando como argumento de entrada al filtro *imtophat*. Ésta es una técnica de mejora de contraste y transformación de intensidad.

Figura 18. *Imagen original, filtrado top-hat y mejora de visibilidad con imadjust*



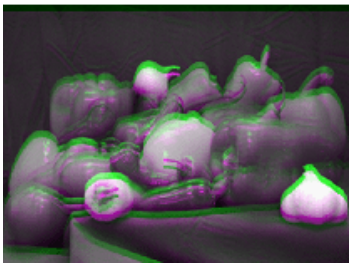
Obtenido de [58].

7.3. Segmentación

La fase de la segmentación se basa en crear imágenes binarias a partir de un valor umbral y las regiones médicas de interés. Por ende, se limpian los bordes de la imagen binaria obtenida de la mejora del contraste mediante la ecualización de histograma previa, resaltando las áreas internas y eliminando toda región conectada a los márgenes. Esto se lleva a cabo a partir del comando *imclearborder*, que suprime estructuras en la imagen que son más claras que los entornos y que están conectadas con el borde de la imagen. Además, debido a que se cuenta con el informe condensado con datos específicos para cada uno de los casos, se hace uso de las regiones de interés (ROI). El procesamiento basado en ROI se representa como una máscara binaria con forma y posición específica, por lo que a partir de las coordenadas *(x, y)* conocidas del centro de la anomalía y el radio aproximado (en píxeles) de la lesión; se procede a graficarlo para validar la capacidad del algoritmo. Del mismo modo,

se realizan comparaciones entre las imágenes a través de la función *imshowpair*, con método de visualización *falsecolor* (Figura 19) y color de salida *green-magenta*. En consecuencia, se determina que las regiones verdes indican dónde tienen las imágenes la misma intensidad, mientras que las regiones coloreadas de magenta muestran dónde difieren las intensidades.

Figura 19. *Método de visualización falsecolor a través de comando imshowpair*

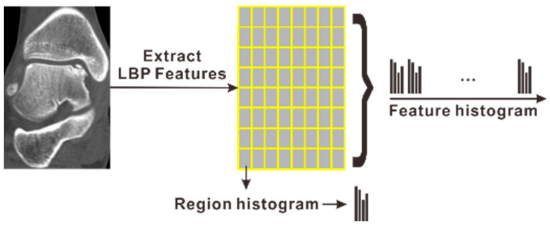


Nota. Se superponen dos bandas de color diferentes. Las regiones grises indican dónde las imágenes tienen la misma intensidad. Las regiones coloreadas muestran dónde difieren las intensidades. Obtenido de [59].

7.4. Posprocesamiento

Como parte de la fase de posprocesamiento, debido a que las visualizaciones obtenidas a partir del comando *imshowpair* implica un canal triple de color de salida, es necesario extraer la información de cada uno de los canales de las dos imágenes que componen la superposición. Se extraen las características del patrón binario local (LBP) que codifican información de textura local (*extractLBPFeatures*) siendo útil para indicar si existe o no presencia de algún tumor, y se realiza un recuento de bins del histograma con una cantidad de 256 niveles de gris que pueden tener las imágenes en formato de 8 bits (*histcounts*) (Figura 20). El conteo del número de elementos en cada uno de los bins o intervalos definidos dentro del rango de los datos, permite analizar la distribución de intensidades en las imágenes, lo cual es útil para la extracción de características, la identificación de los distintos tipos de tejidos mamarios y cálculo de la densidad mamaria.

Figura 20. *Extracción de las características del patrón binario local y recuento de intervalos de intensidad en imagen médica*



Obtenido de [60].

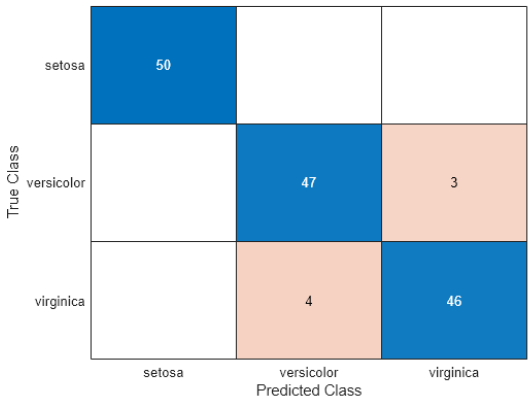
Además, también se extraen las características de forma e intensidad; con el propósito que el modelo a entrenar contenga mayor información y patrones que le permitan aumentar su precisión y robustez. Entre las características de forma se encuentra el área, perímetro, excentricidad y solidez. Dichas características se extraen a través del comando *regionprops*, una función que mide las propiedades de cada uno de los objetos (componentes conectados) de una imagen binaria. En cuanto a las características de intensidad, se encuentran las medias y desviación estándar, que proporcionan información con respecto a la densidad del tejido y la homogeneidad de las áreas de interés a través de los comandos *mean* y *std*.

7.5. Clasificación

Para las 322 mamografías, se procede a convertir las etiquetas categóricas correspondientes a la gravedad de la anomalía (B - Benigno, M - Maligno) en índices numéricos, siendo 1 para B - Benigno y 2 para M - Maligno. Esta asignación de un número entero único a cada categoría facilita el manejo de etiquetas en el algoritmo de aprendizaje automático por medio de *machine learning*. Luego, se realiza una partición de los datos *cvpartition* utilizando el método *'HoldOut'*, es decir que el 20 % de los datos se destina para pruebas, mientras que el 80 % se utiliza para entrenamiento; lo que permite evaluar la capacidad del modelo. Cabe destacar que al dividir los 322 estudios mamográficos en dos conjuntos, se entrena el modelo y se evalúa el rendimiento en base a las características extraídas tanto de textura local como de la distribución de intensidades y características de forma, dando como resultado una estimación realista de cómo el modelo funcionaría en datos no conocidos ni vistos, además de evitar el sobreajuste.

Se entrena un modelo de clasificación basado en una máquina de vectores de soporte (SVM) *fitcsvm* utilizando el conjunto de datos de entrenamiento. A partir del modelo entrenado se efectúan predicciones sobre el conjunto de datos de prueba, calculando la proporción de predicciones correctas, que representa la precisión del modelo. Se crea una matriz de confusión a partir del comando *confusionchart*, que compara las etiquetas verdaderas con las predicciones, mostrando tanto las veces que el modelo predijo cada clase correctamente, como las veces que confundió una clase con otra. Por ende, se delimita en este ejemplo que tres mediciones de la clase 'versicolor' están mal clasificadas, mientras que todas las mediciones pertenecientes a 'setosa' y 'virginica' están clasificadas correctamente (Figura 21).

Figura 21. *Matriz de confusión para un conjunto de datos del iris de Fisher*



Obtenido de [61].

Finalmente, se evalúa el modelo creado pero con nuevas mamografías que no son parte del conjunto de entrenamiento o prueba. Para ello, se recurre a la alianza con el Hospital El Pilar, con el propósito de obtener estudios mamográficos anonimizados para evaluar la precisión del modelo con datos desconocidos; además de conocer la perspectiva del Dr. Dardón, en cuanto al modelo y a las imágenes procesadas obtenidas a partir del flujo de trabajo.

CAPÍTULO 8

Resultados

Las imágenes médicas digitales son propensas a varios tipos de ruido como resultado de errores en el proceso de adquisición. En consecuencia, los valores de los píxeles de las imágenes no reflejan las verdaderas intensidades y por ende, requieren de la aplicación de un flujo de trabajo que inicia con la importación y preprocesamiento, para luego efectuar una segmentación, posprocesamiento y con ello, lograr una clasificación correcta de masas.

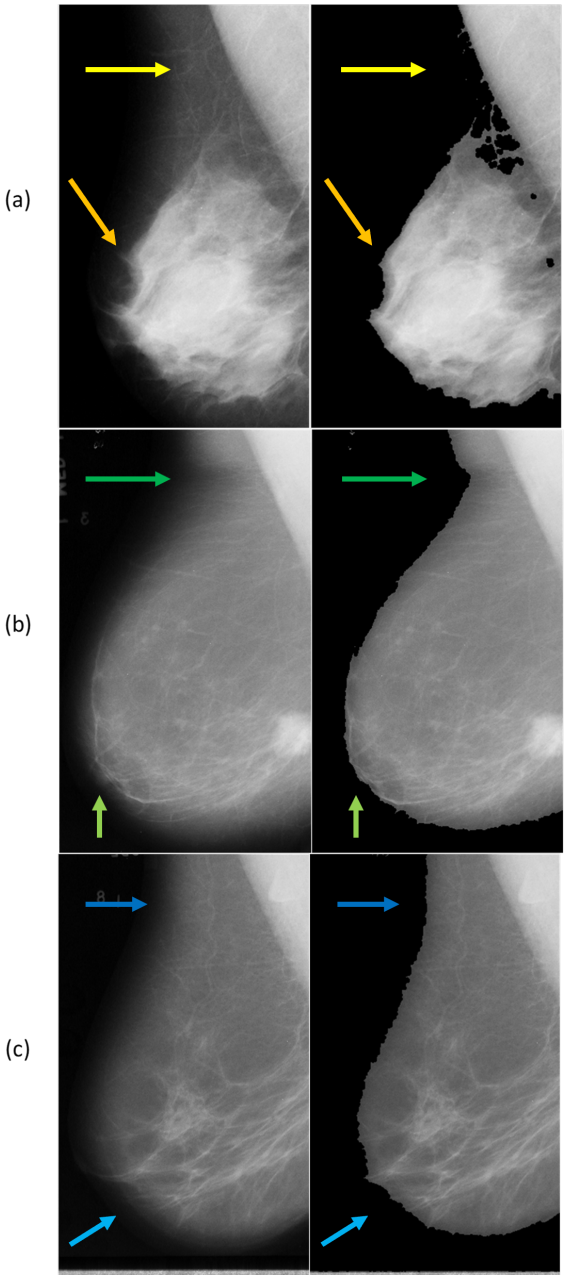
8.1. Importación

Por ende, en función del flujo de trabajo aplicado 322 mamografías y la automatización que itera sobre cada uno de los archivos con formato .pgm, se logra la extracción de información complementaria para cada estudio mamográfico; siendo: gravedad de anomalía (B - Benigno / M - Maligno), coordenadas x, y (centro de la anomalía) y radio aproximado (círculo en píxeles que encierra la anomalía).

8.2. Preprocesamiento

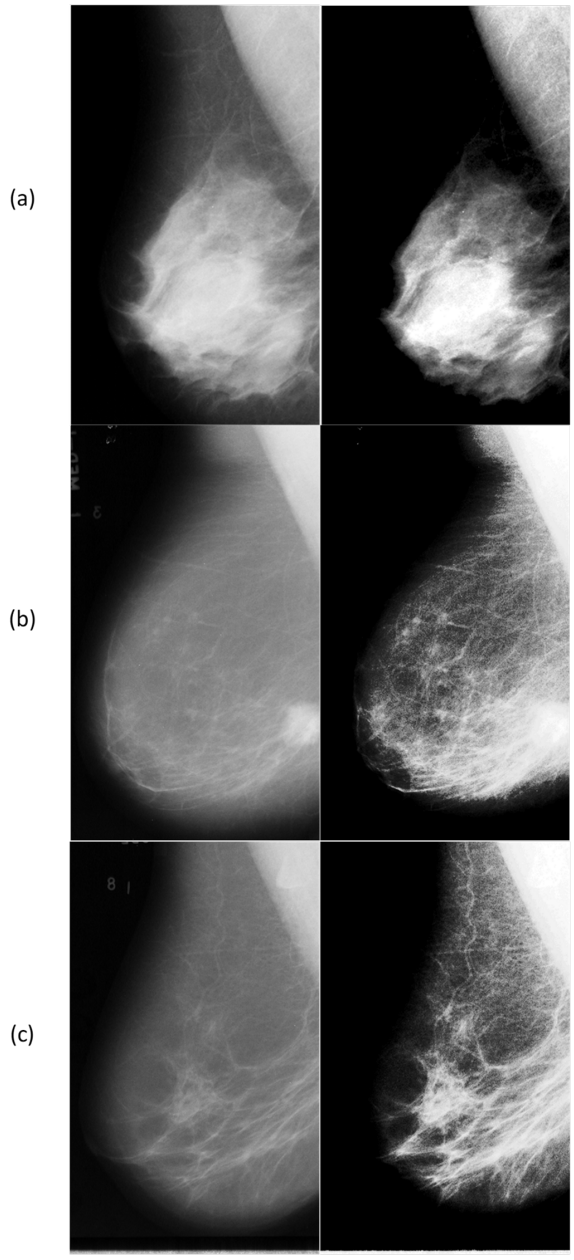
Tras aplicar el filtro de mediana *medfilt2*, la binarización *imbinarize* y el cierre morfológico *imclose* al set de mamografías, es evidente que se obtuvieron imágenes homogeneizadas, estandarizadas y de mayor calidad (Figura 22). Cabe mencionar que para fines visuales explicativos, solamente se presentan 3 de 322 pares de mamografías, del lado izquierdo se muestran las mamografías crudas, mientras que del lado derecho se evidencia la imagen médica luego del preprocesamiento. A partir de ello, es evidente que tanto para la mamografía de tipo benigno, como maligno y normal, se manifiesta una menor distorsión, una mejora de textura y bordes circunscritos.

Figura 22. *Mamografías con filtros de preprocesamiento aplicados, siendo de tipo (a) benigno, (b) maligno y (c) normal. Bordes más definidos*



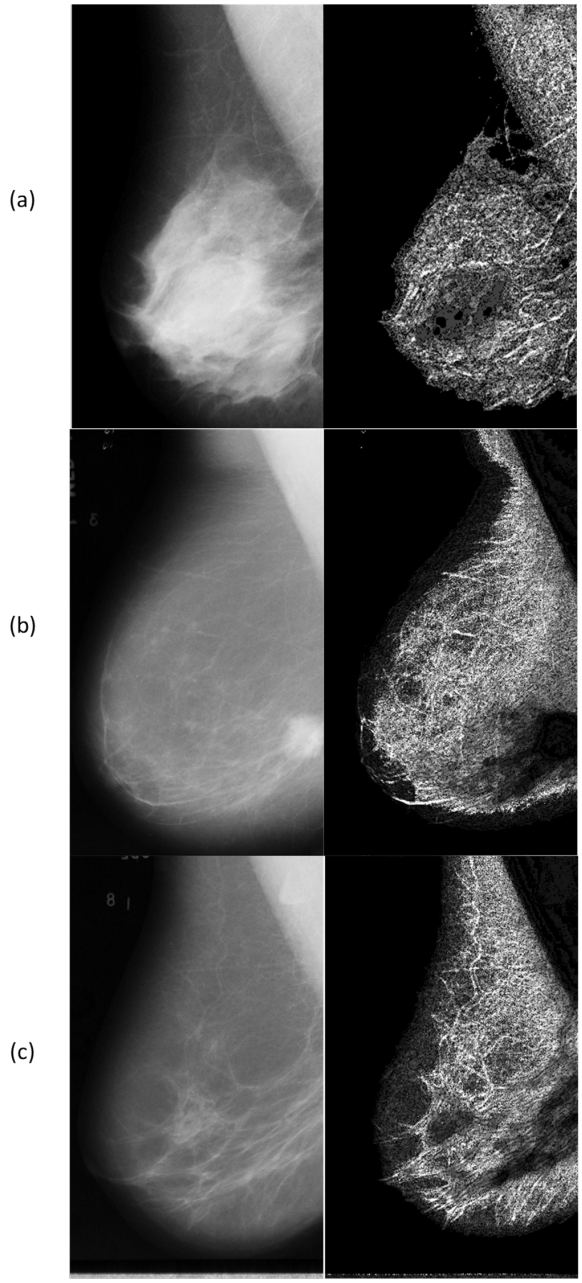
Al aplicar la ecualización del histograma *histeq* para cada una de las mamografías, se mejoró el contraste de las mismas, pero sin perder información de tipo estructural. Además, se obtuvieron imágenes médicas con valores de píxeles distribuidos de forma uniforme, lo que permite observar tejidos de distintas densidades y un realce de zonas correspondientes a tejido mamario y fibroglandular (Figura 23).

Figura 23. *Ecualización de histograma a mamografía (a) benigna, (b) maligna y (c) normal. Distinción entre tejido fibroglandular, estructural y graso de la mama*



Luego, tras aplicar el filtrado *top-hat* y el ajuste de los valores de intensidad *imadjust*, se permitió una corrección de la iluminación, intensidad y contraste de las mamografías. Se logró un resalte de todo aquel detalle o estructura clara en presencia de fondos sombríos (Figura 24), que habían sido eliminados en el proceso de cierre aplicado anteriormente.

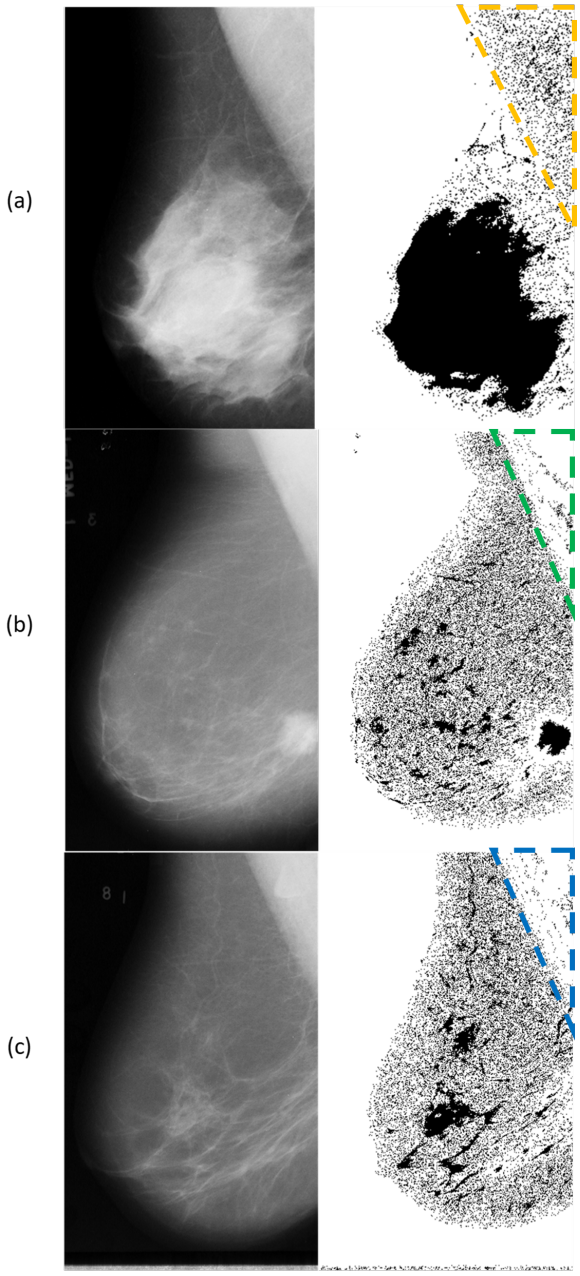
Figura 24. *Aplicación de imtophat para el filtrado de objetos pequeños por el píxel vecino más próximo, para los tipos de mamografías (a) benigna, (b) maligna y (c) normal*



8.3. Segmentación

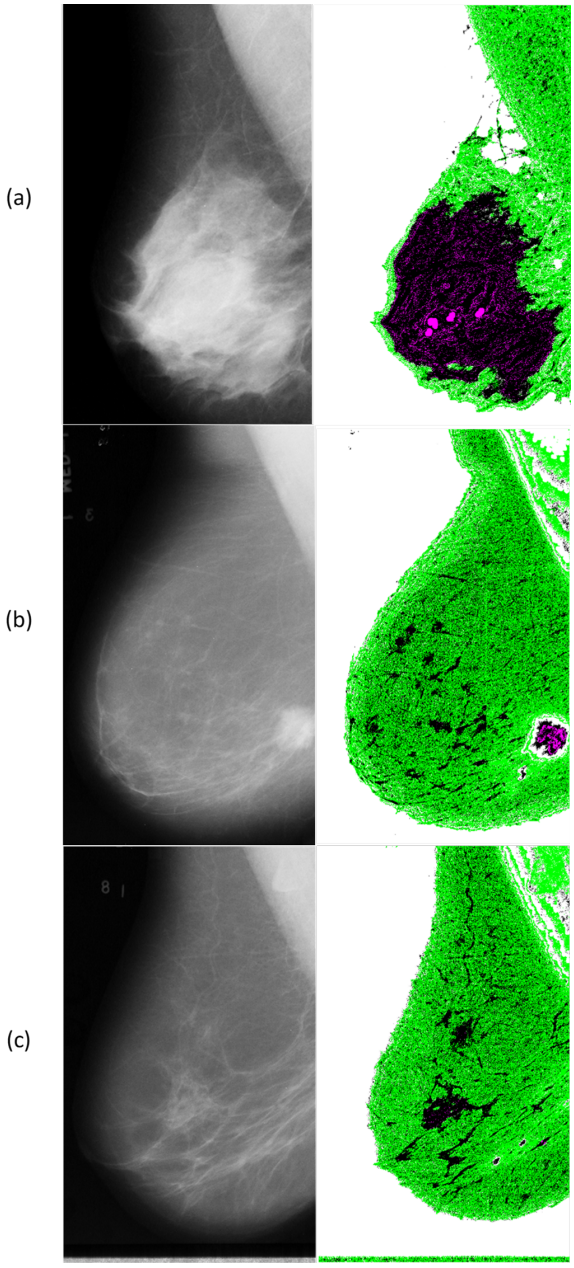
Por otra parte, la función *imclearborder* permitió suprimir el músculo pectoral con un aspecto más claro que el entorno de la mama y que se encontraba conectado con el borde de la imagen mamográfica. Se obtuvieron mamografías sin las estructuras de los bordes correspondientes a los músculos que conectan la parte delantera del pecho humano con los huesos de la parte superior del brazo y el hombro (Figura 25).

Figura 25. *Eliminación de las estructuras de línea conectadas al borde de una mamografía (a) benigna, (b) maligna y (c) normal. Segmentación del músculo pectoral*



Al hacer uso de las regiones de interés (ROI) se logran graficar las masas en las coordenadas dadas por el informe de la base de datos; lo que permite realizar comparaciones con las lesiones encontradas experimentalmente tras aplicar el flujo de trabajo antes mencionado. Como resultado, para cada mamografía se obtiene una imagen compuesta por dos regiones, siendo verde dónde ambas imágenes tienen la misma intensidad y magenta para las regiones donde difieren las intensidades (Figura 26).

Figura 26. *Segmentación de mama y lesión para caso (a) benigno, (b) maligno y (c) normal. Color magenta áreas con lesiones y color verde regiones sanas de la mama*



8.4. Posprocesamiento

Como resultado de la imágenes compuestas por dos regiones, se logra la extracción de características de ambos canales. Debido a que las áreas condensadas de magenta se asocian a la presencia de tumores o anormalidades; el canal magenta destaca las regiones en donde existe un aumento de la densidad del tejido mamario; suponiendo la presencia de alguna lesión. Mientras que el componente verde, representa el fondo de la mama o la región sana.

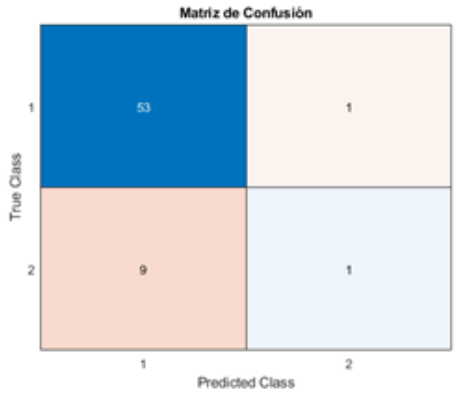
Tras la extracción de las características de textura *extractLBPFeatures*, se logra una captura de texturas locales en cada mamografía, siendo útil para indicar si existe o no la presencia de algún tumor. Por el contrario, un recuento de bins del histograma *histcounts* permite representar la distribución de las intensidades de color y con ello, identificar los distintos tipos de tejidos mamarios.

Adicionalmente a las dos características antes mencionadas, también se extraen 4 características adicionales que corresponden a la forma, 1 a la intensidad y 1 a la binarización. Las características de forma incluyen: área, perímetro, excentricidad y solidez. El área es la extensión en píxeles, siendo un área inusualmente grande signo de posible presencia de tumor. El perímetro permite determinar la irregularidad de la forma del tumor, mientras que la excentricidad delimita la forma del tumor. La característica de intensidad corresponde a las medias y desviación estándar, ambas proporcionan información con respecto a la densidad del tejido y la homogeneidad de las áreas de interés. Finalmente, la característica de binarización proporciona una representación clara de las áreas de interés (tumores) y el fondo.

8.5. Clasificación

En conjunto, todas estas características conforman patrones de suma relevancia para el entrenamiento del modelo de clasificación basado en máquina de vectores de soporte (SVM) *fitcsvm*. Por ende, tras una partición de las 322 mamografías, siendo el 80 % para entrenamiento y el 20 % para prueba; se logra la clasificación de nuevos datos en función de las características aprendidas con una precisión del 84 % para predecir las etiquetas de nuevos casos que son parte del conjunto de prueba. La matriz de confusión (Figura 27) muestra el número total de observaciones en cada celda, donde las filas corresponden a la clase verdadera y las columnas a la clase predicha. Por lo tanto, las celdas diagonales y no diagonales corresponden a las observaciones clasificadas correctamente e incorrectamente, respectivamente. A partir de ello, con una distribución de clases en el conjunto de entrenamiento de: 217 y 41, y una distribución de clases en el conjunto de prueba de 54 y 10; el conjunto de prueba permitió una clasificación de 53 verdaderos positivos, 1 verdadero negativo, 9 falsos positivos y 1 falso negativo.

Figura 27. *Matriz de confusión del modelo machine learning, aprendizaje supervisado por medio de máquina de vectores de soporte (SVM)*



En función del modelo entrenado y con un set de 10 mamografías anonimizadas de pacientes guatemaltecas, se logra en su mayoría la detección de las lesiones establecidas por los radiólogos. Sin embargo, aún se cuenta con ciertas limitantes para el caso de masas muy pequeñas. Sin embargo, en cuanto a la retroalimentación obtenida por parte del Dr. Aldo Dardón del Hospital El Pilar, se determina que en general el algoritmo es eficiente y preciso; con un alto alcance en mamas mayormente conformadas por tejido fibroglandular y patrones densos.

CAPÍTULO 9

Discusión

9.1. Importación y preprocesamiento

Las mamografías se sometieron a distintos filtros y técnicas con la intención de mejorar los datos de la imagen para que pudieran ser utilizados para su posterior procesamiento (Figura 22) [7]. La función *medfilt2* constituyó una de las herramientas esenciales debido a su capacidad para mejorar la calidad de la imagen y reducir el ruido o cualquier artefacto asociado a la adquisición [54] [63], en este caso de la mamografía o compresión del tejido mamario. Por lo tanto, este filtrado fue eficaz para eliminar el ruido de impulso que suele producirse durante la digitalización de la imagen, dando un resultado satisfactorio en el filtrado de las imágenes para eliminar el ruido innecesario [63] [64]. Además, es evidente que el filtro de mediana bidimensional suavizó la mama, al mismo tiempo que aumentó el contraste global, corrigió distorsiones geométricas y mejoró el contraste local de bordes, tejidos y masas.

En cuanto a la binarización *imbinarize*, se puede destacar que se identificaron los límites de la imagen y se separaron los píxeles según el umbral [7] [55] [65]. Por ende, se estableció que las áreas con mayor densidad corresponden a tumores o masas; esto debido a que los píxeles de las mamografías presentaban diferentes niveles de intensidad (umbral) que representan diferentes tonos de gris y diferentes áreas del seno. Es por ello que, la binarización resaltó las áreas de interés (calcificaciones) al separarlas del fondo, facilitando su identificación con respecto a los tejidos circundantes [62].

La función *histeq* se utilizó para realizar ajustes de la distribución de los niveles de intensidad, un proceso que mejoró el contraste visual y la visibilidad de las características relevantes [57]. Esto es importante porque una visualización clara de las áreas de interés aumenta la precisión en la detección de anomalías. En mamografías, las microcalcificaciones son pequeños depósitos de calcio en el tejido mamario y son uno de los primeros signos de

cáncer de mama [19] [21]. Sin embargo, suelen ser difíciles de distinguir si el contraste entre ellas y el tejido circundante es bajo [1]. Por tanto, la ecualización del histograma ayudó a hacer estas características más visibles y prominentes, mejorando el contraste local [62], dado que los tejidos cancerosos son muy blancos, el valor más alto de los píxeles es el interés principal [65]. Además, a veces las mamografías contaban con áreas con una iluminación desigual debido a factores como artefactos durante la captura o la compresión del tejido. Entonces, es evidente que *histeq* ayudó a nivelar la distribución de la intensidad de luz en la imagen, ya que se reducieron las zonas oscuras o brillantes y se obtuvieron imágenes más homogéneas, lo que facilitó un análisis más uniforme (Figura 23). La imagen se mejoró mediante el estiramiento del contraste que ajustó el histograma de la imagen para que hubiera una mayor separación entre la distribución del nivel de grises del primer plano y del fondo. El contraste y el brillo son muy importantes para mejorar las imágenes de la mamografía, ya que se trata de imágenes binarias, por lo tanto, cuanto más contraste entre la partícula blanca y la partícula negra, más claras se pueden ver las partículas blancas [65].

Por otra parte, el filtrado *top-hat* resaltó los detalles pequeños y finos de la imagen. Esta técnica es particularmente útil cuando se desea extraer detalles que pueden ser difíciles de identificar debido a un fondo uniforme o de bajo contraste [58], específicamente en las microcalcificaciones. Las microcalcificaciones son típicamente de baja intensidad y pueden ser difíciles de identificar debido al ruido o al fondo heterogéneo, por lo que el filtrado *top-hat* resaltó las pequeñas características brillantes que correspondían a pequeñas acumulaciones de calcio. Además, en las mamografías las áreas de tejido denso o las sombras pueden oscurecer o enmascarar detalles importantes, por ende el filtrado *top-hat* ayudó a eliminar la información del fondo y resaltó las estructuras relevantes, lo que facilitó la detección de anomalías [65]. La función *imadjust* realzó el contraste en áreas con bajo contraste, lo que mejoró la claridad de nódulos o masas, destacó detalles sutiles (como bordes de masas o pequeñas calcificaciones), ajustó las variaciones ocasionadas por la iluminación debido a la compresión del tejido o el ángulo de toma y mejoró la visibilidad y uniformidad en toda la mama (Figura 24).

9.2. Segmentación

La segmentación de imágenes se utilizó para dividir una imagen en partes que tenían características y propiedades similares, simplificándola de forma fácilmente analizable [47] [65]. La función *imclearborder* suprimió estructuras en la imagen que eran más claras que los entornos y que estaban conectadas con el borde de la imagen [66]. Además, permitió eliminar el músculo pectoral, el cual suele presentar un contraste más claro y brillante en comparación con el tejido mamario y está conectado con los bordes de la imagen [64]. Por ende, a través de *imclearborder* se logró reducir la posibilidad de que el proceso de análisis se viera afectado por irregularidades, se segmentó el diagnóstico solamente a la región de interés y no a las estructuras periféricas, y se facilitó la detección de anomalías en las áreas de interés (Figura 25).

Con la aplicación de *imshowpair* se observaron dos tipos de regiones. Las regiones verde correspondían a las áreas en donde la intensidad era la misma, mientras que las áreas de color magenta mostraban las zonas en las que la intensidad difería [59]. Por lo tanto, se

logró una visualización de masas o lesiones como áreas concentradas o con alto contraste en color magenta, mientras que en color verde se mostraron las regiones que correspondían a la anatomía y tejidos asociados de la mama. En general se estableció que las áreas magenta se asociaban a regiones con aumento de la densidad del tejido mamario suponiendo la presencia de tumores o anormalidades; mientras que el componente verde, representaba el fondo de la mama o la región sana (Figura 26). Añadiendo las regiones de interés (ROI), se validó la presencia de las lesiones [67]; y al marcar las ubicaciones de las calcificaciones y compararlas con las áreas encontradas experimentalmente, se facilitó una evaluación visual y se estableció una concordancia entre las lesiones reales descritas por el informe del *dataset* y las predicciones del modelo. En definitiva, las imágenes resultantes recopiladas de este proceso pueden ser útiles para ayudar a los radiólogos a un diagnóstico de cáncer de mama y monitoreo del proceso de tratamiento [7].

Sin embargo en algunos casos fue evidente el enmascaramiento, ya que se hallaron lesiones adicionales no establecidas en el reporte de la base de datos. Las lesiones adicionales detectadas representaron y supusieron una ventaja significativa en el diagnóstico de calcificaciones a partir de mamografías, ya que implicaba que el modelo podría potencialmente identificar masas que pasaron por desapercibidas. Por ende, las lesiones no reportadas plantearon la cuestión de la calidad del *dataset* y la uniformidad en el etiquetado de las imágenes en cuanto a la distribución de calcificaciones. De acuerdo al reporte, en muchos casos en donde existían varias calcificaciones, las ubicaciones centrales y los radios se aplicaron solamente a los grupos en lugar de a calcificaciones individuales; mientras que en otros casos, las ubicaciones centrales y los radios se omitieron para las calcificaciones que se encontraban ampliamente distribuidas por toda la imagen en lugar de estar concentradas en un solo sitio [53]. Por lo tanto, se determinó que el enmascaramiento de las lesiones implicó una limitación de la base de datos, un reflejo de las dificultades inherentes humanas para la segmentación de estructuras complejas y una baja resolución de las imágenes mamográficas que afectó la calidad de las mismas; suponiendo una debilidad para identificar pequeñas calcificaciones efectivamente. Otra causa del enmascaramiento también se asoció a los senos densos, ya que cuando las mamas son muy densas la mamografía tiene sus limitaciones con una sensibilidad de aproximadamente el 70 %, provocando que los cánceres puedan estar ocultos en los estudios mamográficos [1].

En consecuencia, un sistema capaz de detectar lesiones no observadas inicialmente podría mejorar significativamente la detección temprana de cáncer de mama, identificando tumores o calcificaciones que podrían haber sido pasados por alto por radiólogos humanos, especialmente en casos de calcificaciones difusas, mamas densas o lesiones pequeñas. Sin embargo, se consideró que la detección de lesiones adicionales debe ser validada cuidadosamente para eludir interpretaciones erróneas y falsos positivos. En este contexto, la colaboración entre sistemas automáticos y expertos médicos es fundamental para asegurar que las detecciones del modelo sean precisas y útiles en la práctica clínica médica.

9.3. Posprocesamiento

Las lesiones mamarias se pueden caracterizar como benignas o malignas en función de su forma, textura y valores de intensidad del nivel de gris [64]. Por lo tanto, con el fin de

extraer estas las características o *features* de textura de ambos canales (magenta y verde), se realizó un recuento de bins del histograma *histcounts*, lo que permitió representar la distribución de las intensidades de color [68] y con ello, identificar los distintos tipos de tejidos mamarios, así como determinar la presencia o no de lesiones a partir de patrones encontrados en el análisis de las 322 mamografías. Adicionalmente las características de forma (área, perímetro, excentricidad y solidez) admitieron la capacidad de evaluar la irregularidad y forma del tumor, extensión en píxeles que representa el área inusual, densidad del tejido y homogeneidad e irregularidad de las áreas de interés y de sus márgenes [64].

El perímetro evaluó la irregularidad de la mama que puede ser un indicador clave de que existe una masa sospechosa o tumoral. Los tumores, en especial los malignos, suelen tener contornos irregulares y no definidos, lo que puede diferenciarse de las formaciones benignas o del tejido mamario normal; como indicador clave del parámetro de excentricidad [24] [29]. La solidez determinó cuán sólido es la masa y proporcionó información sobre el tipo de tumor, debido a que un tumor menos sólido puede ser más difuso o menos compacto, lo que puede ser indicativo de una lesión más agresiva.

De ahí que, las características extraídas de los ROI también son importantes para mejorar el rendimiento del clasificador; siendo las de textura y forma las más importantes para distinguir las lesiones benignas de las malignas [64]. Así pues, el conjunto de características facilitó la clasificación o la predicción de condiciones patológicas para el modelo de *machine learning*; ya que el análisis de un conjunto de atributos repetitivos se asocia con una mejora del proceso de discriminación entre tejidos normales y patológicos mamarios. Además, permitió al modelo de clasificación discernir entre tejidos mamarios normales y anómalos a partir de la homogeneidad y heterogeneidad de la mamografía y una reducción de falsos positivos y falsos negativos al centrar la atención en las áreas relevantes.

9.4. Clasificación

La máquina de vectores de soporte (SVM) es un método de aprendizaje estadístico supervisado y automático eficaz para la clasificación [5], con altas tasas de distribución en el cáncer de mama, ordenando los datos en categorías [65]. Entonces la clasificación de las mamografías utilizando el modelo SVM, evidenció un desempeño bueno con una precisión del 84 % (Figura 27); lo que implica una capacidad alta para predecir las etiquetas y clasificar en benigno o maligno nuevos casos desconocidos. Además, la métrica de la precisión estableció que el modelo ha aprendido eficazmente a partir de las características relevantes extraídas con la técnica *LBP* y que el flujo de trabajo de importación, preprocesamiento, segmentación y posprocesamiento aplicado permite detectar lesiones mamarias a partir de estudios mamográficos.

Por otra parte, el modelo destacó gran capacidad para identificar correctamente los casos positivos, con 53 verdaderos positivos en un total de 54 casos en el conjunto de prueba; lo que sugirió alta tasa de detección de lesiones mamarias. Dicho desempeño es relevante, ya que en la detección temprana del cáncer de mama, un alto número de verdaderos positivos indica que el modelo está detectando correctamente la mayoría de los tumores, lo que puede tener un impacto directo en la mejora de los resultados, práctica y diagnósticos clínicos; ayudando en la intervención temprana contra el mal que aqueja a muchas mujeres.

Además, se contó con 1 caso correctamente clasificado como negativo (sin tumor o anomalía), 1 caso incorrectamente clasificado como negativo cuando debería haber sido positivo y 9 casos incorrectamente clasificados como positivos cuando deberían haber sido negativos. Estos últimos 9 casos representaron los falsos positivos que podrían conllevar a la realización innecesaria de procedimientos adicionales o complementarios, como biopsias o pruebas más invasivas. Si bien los falsos positivos no son deseables, tienen un impacto clínico menos grave en comparación con los falsos negativos, ya que los primeros suelen generar un exceso de precaución. Por el contrario, el único caso falso negativo podría llegar a tener consecuencias graves, ya que es capaz de retrasar el diagnóstico y tratamiento de un caso de cáncer de mama. Por ende, la minimización de falsos negativos es crucial, y en este caso, el modelo muestra un buen rendimiento, pero aún tiene un pequeño margen de mejora. En base a ello, se consideró que se pueden realizar mejoras en la segmentación del músculo pectoral y en la densidad propia de la mama, ya que ambas variables implican pequeñas variaciones en cuanto a las características extraídas (patrones) y a la proporción de tejido denso y no denso que implican una clasificación al siguiente nivel.

En cuanto al procesamiento de las 10 mamografías pertenecientes a mujeres guatemaltecas, se establece que el algoritmo es capaz de segmentar correctamente las regiones con masas y las áreas sanas de la mama; dando paso en su mayoría a una clasificación precisa. Cabe destacar que en muchos casos donde la mama es muy densa, el tejido fibroglandular impide la visualización del tumor a través de una inspección puramente visual, subjetiva y cualitativa; y por ende, se requieren de estudios complementarios como ultrasonidos [22]. En una mamografía, el tejido mamario graso es transparente y por ende, es fácil ver a través de este para detectar alguna masa. Por el contrario, el tejido mamario denso tiene apariencia blanca y sólida, siendo difícil ver calcificaciones a través de este debido a que los quistes o lesiones también tienen una apariencia blanca y sólida [23]. Además es importante tomar en cuenta que, la densidad del tejido mamario se asocia con un mayor riesgo de cáncer de mama y dificulta la obtención de resultados de los exámenes de detección [34]. Por ende, el algoritmo supone ser una herramienta de apoyo para la detección temprana de cánceres ocultos o en estadios precoces, en especial en los casos donde la densidad mamaria es un factor de riesgo o cuando la capacidad económica del paciente se encuentra restringida para realizarse estudios de imágenes médicas adicionales.

En general todos los casos se segregaron correctamente entre benigno y maligno. Además, se estableció que el flujo de trabajo aplicado a las mamografías permite una detección correcta de casos benignos y malignos, ya que es evidente la segmentación entre tejido denso glandular y lesiones anormales. Por ende, se logró la obtención imágenes médicas con mayor detalle, siendo posible diferenciar a simple vista los tejidos propios del seno; así como signos de calcificaciones, bultos, crecimientos inusuales o tumores con texturas, formas, umbrales de color, máscaras y distribución de tejido diferentes al resto de la mama.

CAPÍTULO 10

Conclusiones

- El flujo de trabajo conformado por las fases de importación, preprocesamiento, segmentación y posprocesamiento permiten obtener imágenes mamográficas más estandarizadas; siendo evidente estructuras correspondientes a la anatomía, histología y morfología propia de la mama que no es posible apreciar a través de una inspección puramente visual. Sin embargo, esto depende en gran medida de la calidad de la mamografía cruda, ya que muchas de éstas representan variaciones causadas por artefactos provenientes de la adquisición, por el equipo (mamógrafo), por la compresión aplicada al seno o por la variabilidad y complejidad de las anormalidades.

- La densidad mamaria constituye un aspecto influyente que afecta en la segmentación y detección de lesiones con respecto a estructuras periféricas del seno. Asimismo, el músculo pectoral repercute en el proceso de segmentación y extracción de características, ya que su constitución histológica representa una posible área de interés con una apariencia similar a las zonas con anomalías. Cabe mencionar que también se presenta dificultad para encontrar lesiones pequeñas poco densas.

- El algoritmo es capaz de hallar lesiones adicionales no reportadas por el informe de la base de datos. La causa principal se asocia a calcificaciones que fueron omitidas debido a su amplia distribución o a la presencia de grupos de lesiones que fueron consideradas como calcificaciones individuales. Por ende, se plantea tanto la calidad del *dataset*, como la uniformidad en el etiquetado y clasificación en cuanto a la distribución de las calcificaciones. Sin embargo, también es importante tomar en cuenta que la interpretación y el análisis de las mamografías del *dataset*, corresponde a una clasificación basada en hallazgos y una inspección visual por parte de médicos radiólogos; lo que se vincula a un posible error humano, enmascaramiento de las lesiones a causa de la densidad mamaria y falta de herramientas para una visualización con mayor detalle.

- El modelo de aprendizaje supervisado para la clasificación (SVM) cuenta con una precisión del 84 % a partir del conjunto de prueba. Es capaz de detectar lesiones mamarias, con un número alto de verdaderos positivos, lo que indica su capacidad para detectar correctamente la mayoría de tumores y clasificarlos en las dos clases.

- A pesar de la alta capacidad del modelo para predecir las etiquetas y clasificar los tumores mamarios en benignos y malignos, no se considera una herramienta de diagnóstico médico. Más bien, se cataloga como un algoritmo de apoyo para la detección temprana de cánceres en estadios precoces o cuando la densidad mamaria es un factor de riesgo del paciente. Además, el modelo también se considera como una herramienta de acompañamiento para efectuar un diagnóstico más preciso y eficaz, proporcionando información adicional al radiólogo para que se realice la interpretación médica en conjunto con la experiencia y el análisis de variables no implícitas o desconocidas por parte del modelo.

- En general, todos los casos son etiquetados correctamente en su clasificación correspondiente (maligno o benigno), tomando en cuenta el aprendizaje de los patrones a partir de las *features* extraídas y el flujo de trabajo aplicado. Una pequeña proporción irrelevante de las mamografías reportadas como benignas, fueron clasificadas como malignas y viceversa.

- Se determina que el modelo de ML creado es capaz de predecir correctamente la clasificación de mamografías desconocidas correspondientes a pacientes guatemaltecas, que no son parte del conjunto de prueba ni entrenamiento; además de detectar lesiones ocultas en tejidos mamarios densos conformados en su mayoría por tejido fibroglandular que causan enmascaramiento de lesiones.

CAPÍTULO 11

Recomendaciones

- Se recomienda aplicar técnicas 'clásicas' de procesamiento de imágenes, como lo es *thresholding*; para evaluar la eficacia con respecto a la ecualización de histograma. Además, del uso de apps de Matlab, como *image segmenter*, *color thresholder* o *image region analizer*; para comparar los distintos resultados de hallazgos.
- Se sugiere aplicar *deep learning*, específicamente redes neuronales convolucionales (CNN) tomando en cuenta el uso de las funciones de activación ReLU, que provocan que las redes profundas sean más expresivas que sus contrapartes no profundas, aumentando la cantidad de capas y por consiguiente, las regiones lineales.
- Se propone hacer uso de 3D Slicer en conjunto con Matlab, con el fin de mejorar las fases de procesamiento y segmentación. Además, como herramienta complementaria para la planificación y asistencia (guía) cuando es necesario realizar procedimientos invasivos como estudios complementarios (biopsias).
- Se aconseja entrenar un modelo basado en la extracción de características de la mamografía cruda, con el propósito de evaluar si existe una discrepancia de los patrones adquiridos para la clasificación en las dos clases y con ello, determinar las variables más significativas y establecer si aumenta la precisión.
- Se insta a utilizar una base de datos con mayor volumen de mamografías, con mayor información relevante del paciente y variables asociadas, para que el algoritmo sea capaz de clasificar en base a características adicionales; dando paso a un aprendizaje basado en patrones hallados.
- Se recomienda evaluar una clasificación multiclase según el sistema Breast Imaging Reporting And Data System (BI-RADS), el cual clasifica los resultados de las pruebas según 1 de 7 categorías, que van desde un resultado normal o benigno (no canceroso) hasta altamente sospechoso o maligno (cáncer). Dicho registro permitirá homogeneizar el método para clasificar los hallazgos mamográficos, categorizando las mamografías en el idioma universal actual empleado para el diagnóstico de la patología mamaria.

CAPÍTULO 12

Bibliografía

[1] J. Hodler, R. A. Kubik-Huch y G. K. von Schulthess, "Diseases of the chest, breast, heart and vessels 2019-2022: diagnostic and interventional imaging," 2019.

[2] E. D. Pisano et al., "Diagnostic performance of digital versus film mammography for breast-cancer screening," *New England Journal of Medicine*, vol. 353, n.º 17, págs. 1773-1783, 2005.

[3] F. Sardanelli et al., "Magnetic resonance imaging of the breast: recommendations from the EUSOMA working group," *European journal of cancer*, vol. 46, n.º 8, págs. 1296-1316, 2010.

[4] M. Puttagunta y S. Ravi, "Medical image analysis based on deep learning approach," *Multimedia tools and applications*, vol. 80, n.º 16, págs. 24365-24398, 2021.

[5] A. Mert, N. Kılıç, E. Bilgili, A. Akan et al., "Breast cancer detection with reduced feature set," *Computational and mathematical methods in medicine*, vol. 2015, 2015.

[6] B. Sahiner et al., "Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images," *IEEE transactions on Medical Imaging*, vol. 15, n.º 5, págs. 598-610, 1996.

[7] V. Patil, S. Burud, G. Pawar, T. Rayajadhav y S. B. Hebbale, "Breast cancer detection using MATLAB functions," *Advancement in Image Processing and Pattern Recognition*, vol. 3, n.º 2, págs. 1-6, 2020.

[8] G. M. Papamija Manzano y J. J. Piamba Muelas, "Desarrollo de una aplicación software para la caracterización BIRADS ecográfica automatizada de lesiones en phantom de mama.," 2021.

[9] J. Tang, R. M. Rangayyan, J. Xu, I. El Naqa e Y. Yang, "Computer-aided detection and diagnosis of breast cancer with mammography: recent advances," *IEEE transactions on information technology in biomedicine*, vol. 13, n.º 2, págs. 236-251, 2009.

[10] O. M. de la Salud, *Cáncer de mama*, 2024. dirección: `https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer#:~:text=En%202020%2C%20685%20000%20personas,todos%20los%20pa%C3%ADses%20del%20mundo`.

[11] O. P. de la Salud, *Cáncer de mama*.

[12] M. Villeda, *Registro Hospitalario 2020*, mayo de 2023. dirección: https://34cdd47e-6421-47cf-8cd6-560fce0dda4a.filesusr.com/ugd/c472b0_104a30deda544355a0%20d8b1163068c309.pdf.

[13] A. Kihn-Alarcón et al., *Breast cancer in young women in Guatemala: A retrospective comparative cohort study.* 2022.

[14] F. Mayo, *Biopsia mamaria*, ago. de 2023. dirección: https://www.mayoclinic.org/es/tests-procedures/breast-biopsy/about/pac-20384812.

[15] X. Liu, J. Liu, D. Zhou y J. Tang, "A benign and malignant mass classification algorithm based on an improved level set segmentation and texture feature analysis," en *2010 4th International Conference on Bioinformatics and Biomedical Engineering*, IEEE, 2010, págs. 1-4.

[16] M. Mehmood et al., "Machine learning enabled early detection of breast cancer by structural analysis of mammograms," *Comput. Mater. Contin*, vol. 67, n.º 1, págs. 641-657, 2021.

[17] B. Gayathri, C. Sumathi y T. Santhanam, "Breast cancer diagnosis using machine learning algorithms-a survey," *International Journal of Distributed and Parallel Systems*, vol. 4, n.º 3, pág. 105, 2013.

[18] O. I. Obaid, M. A. Mohammed, M. K. A. Ghani, A. Mostafa, F. Taha et al., "Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer," *International Journal of Engineering & Technology*, vol. 7, n.º 4.36, págs. 160-166, 2018.

[19] I. N. del Cáncer, *¿Qué es el cáncer?* Mayo de 2021. dirección: https://www.cancer.gov/espanol/cancer/naturaleza/que-es.

[20] N. I. of Health (NIH), *Cáncer metastásico: cuando el cáncer se disemina*, nov. de 2020. dirección: https://www.cancer.gov/espanol/tipos/cancer-metastatico#:~:text=En%20la%20met%C3%A1stasis%2C%20las%20c%C3%A9lulas,en%20otras%20partes%20del%20cuerpo.&text=El%20c%C3%A1ncer%20que%20se%20disemina,lejana%20se%20llama%20c%C3%A1ncer%20metast%C3%A1sico..

[21] A. C. Society, *¿Qué es el cáncer de seno?* Nov. de 2021. dirección: https://www.cancer.org/es/cancer/tipos/cancer-de-seno/acerca/que-es-el-cancer-de-seno.html.

[22] A. C. Society, *Densidad de los senos e informe de su mamograma*, mar. de 2022. dirección: https://www.cancer.org/es/cancer/tipos/cancer-de-seno/pruebas-de-deteccion-y-deteccion-temprana-del-cancer-de-seno/mamogramas/la-densidad-de-los-senos-y-el-informe-de-su-mamograma.html.

[23] "Cáncer de mama," *RECIAMUC*, vol. 6, págs. 521-534, 3 jul. de 2022, ISSN: 25880748. DOI: 10.26820/reciamuc/6.(3).julio.2022.521-534.

[24] L. Choi, *Cáncer de mama*, dic. de 2023. dirección: https://www.msdmanuals.com/es-ec/hogar/salud-femenina/c%C3%A1ncer-de-mama/c%C3%A1ncer-de-mama.

[25] N. O. Bazar, C. B. Hernandez y L. V. Bazar, "Factores de riesgo asociados al cáncer de mama," *Revista Cubana de Medicina General Integral*, vol. 36, n.º 2, págs. 1-13, 2020.

[26] R. G. d. Nascimento y K. M. Otoni, "Histological and molecular classification of breast cancer: what do we know?" *Mastology*, vol. 30, págs. 1-8, abr. de 2020. dirección: https://revistamastology.emnuvens.com.br/revista/article/view/945.

[27] A. C. Society, *Etapas (estadios) del cáncer de seno*, nov. de 2021. dirección: https://www.cancer.org/es/cancer/tipos/cancer-de-seno/comprension-de-un-diagnostico-de-cancer-de-seno/etapas-del-cancer-de-seno.html.

[28] A. C. Society, *Cómo entender su informe de mamograma*, ene. de 2022. dirección: https://www.cancer.org/es/cancer/tipos/cancer-de-seno/pruebas-de-deteccion-y-deteccion-temprana-del-cancer-de-seno/mamogramas/como-entender-su-informe-de-mamograma.html.

[29] S. K. M. Hamouda, R. H. A. E. Ezz y M. E. Wahed, "Enhancement Accuracy of Breast Tumor Diagnosis in Digital Mammograms," *Journal of Biomedical Sciencies*, vol. 06, 04 2017, ISSN: 2254609X. DOI: 10.4172/2254-609X.100072.

[30] E. Arzanova y H. N. Mayrovitz, "The Epidemiology of Breast Cancer," en Exon Publications, ago. de 2022, págs. 1-20. DOI: 10.36255/exon-publications-breast-cancer-epidemiology.

[31] M. Bellanger, N. Zeinomar, P. Tehranifar y M. B. Terry, "Are Global Breast Cancer Incidence and Mortality Patterns Related to Country-Specific Economic Development and Prevention Strategies?" *Journal of Global Oncology*, n.º 4, págs. 1-16, 2018, PMID: 30085889. DOI: 10.1200/JGO.17.00207. eprint: https://doi.org/10.1200/JGO.17.00207. dirección: https://doi.org/10.1200/JGO.17.00207.

[32] B. Smolarz, A. Z. Nowak y H. Romanowicz, "Breast Cancer—Epidemiology, Classification, Pathogenesis and Treatment (Review of Literature)," *Cancers*, vol. 14, n.º 10, 2022, ISSN: 2072-6694. DOI: 10.3390/cancers14102569. dirección: https://www.mdpi.com/2072-6694/14/10/2569.

[33] S. Nardin et al., "Breast Cancer Survivorship, Quality of Life, and Late Toxicities," *Frontiers in Oncology*, vol. 10, 2020, ISSN: 2234-943X. DOI: 10.3389/fonc.2020.00864. dirección: https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2020.00864.

[34] S. Łukasiewicz, M. Czeczelewski, A. Forma, J. Baj, R. Sitarz y A. Stanisławek, "Breast Cancer—Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies—An Updated Review," *Cancers*, vol. 13, n.º 17, 2021, ISSN: 2072-6694. DOI: 10.3390/cancers13174287. dirección: https://www.mdpi.com/2072-6694/13/17/4287.

[35] S. M. Lima, R. D. Kehm y M. B. Terry, "Global breast cancer incidence and mortality trends by region, age-groups, and fertility patterns," *eClinicalMedicine*, vol. 38, pág. 100 985, ago. de 2021, ISSN: 25895370. DOI: 10.1016/j.eclinm.2021.100985.

[36] J. S. Helm y R. A. Rudel, "Adverse outcome pathways for ionizing radiation and breast cancer involve direct and indirect DNA damage, oxidative stress, inflammation, genomic instability, and interaction with hormonal regulation of the breast," *Archives of Toxicology*, vol. 94, págs. 1511-1549, 5 mayo de 2020, ISSN: 0340-5761. DOI: 10.1007/s00204-020-02752-z.

[37] N. I. of Health (NIH), *Tratamiento del cáncer de mama (PDQ®)–Versión para pacientes*, jun. de 2024. dirección: https://www.cancer.gov/espanol/tipos/seno/paciente/tratamiento-seno-pdq.

[38] E. A. Abeelh y Z. AbuAbeileh, "Comparative Effectiveness of Mammography, Ultrasound, and MRI in the Detection of Breast Carcinoma in Dense Breast Tissue: A Systematic Review.," *Cureus*, vol. 16, e59054, 4 abr. de 2024, ISSN: 2168-8184. DOI: 10.7759/cureus.59054.

[39] F. Mayo, *Cáncer de mama*, feb. de 2024. dirección: https://www.mayoclinic.org/es/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475.

[40] J. Seely y T. Alhassan, "Screening for Breast Cancer in 2018—What Should We be Doing Today?" *Current Oncology*, vol. 25, n.º 11, págs. 115-124, 2018, ISSN: 1718-7729. DOI: 10.3747/co.25.3770. dirección: https://www.mdpi.com/1718-7729/25/11/3770.

[41] L. Nicosia et al., "History of Mammography: Analysis of Breast Imaging Diagnostic Achievements over the Last Century," *Healthcare*, vol. 11, n.º 11, 2023, ISSN: 2227-9032. DOI: 10.3390/healthcare11111596. dirección: https://www.mdpi.com/2227-9032/11/11/1596.

[42] N. Fico et al., "Breast Imaging Physics in Mammography (Part I)," *Diagnostics*, vol. 13, n.º 20, 2023, ISSN: 2075-4418. DOI: 10.3390/diagnostics13203227. dirección: https://www.mdpi.com/2075-4418/13/20/3227.

[43] M. Zubair, S. Wang y N. Ali, "Advanced Approaches to Breast Cancer Classification and Diagnosis," *Frontiers in Pharmacology*, vol. 11, 2021, ISSN: 1663-9812. DOI: 10.3389/fphar.2020.632079. dirección: https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2020.632079.

[44] M. S. Jochelson y M. B. I. Lobbes, "Contrast-enhanced Mammography: State of the Art," *Radiology*, vol. 299, n.º 1, págs. 36-48, 2021, PMID: 33650905. DOI: 10.1148/radiol.2021201948. eprint: https://doi.org/10.1148/radiol.2021201948. dirección: https://doi.org/10.1148/radiol.2021201948.

[45] M. Precision, *Mammography*. dirección: https://matsusada.com/application/ps/mammography/.

[46] W. F. Sensakovic et al., "Contrast-enhanced Mammography: How Does It Work?" *RadioGraphics*, vol. 41, n.º 3, págs. 829-839, 2021, PMID: 33835871. DOI: 10.1148/rg.2021200167. eprint: https://doi.org/10.1148/rg.2021200167. dirección: https://doi.org/10.1148/rg.2021200167.

[47] A. E. Burgess, F. L. Jacobson y P. F. Judy, "Human observer detection experiments with mammograms and power-law noise," *Medical Physics*, vol. 28, págs. 419-437, 4 abr. de 2001, ISSN: 0094-2405. DOI: 10.1118/1.1355308.

[48] D. Kopans, S. Gavenonis, E. Halpern y R. Moore, "Calcifications in the Breast and Digital Breast Tomosynthesis," *The Breast Journal*, vol. 17, n.º 6, págs. 638-644, 2011. DOI: https://doi.org/10.1111/j.1524-4741.2011.01152.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1524-4741.2011.01152.x. dirección: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1524-4741.2011.01152.x.

[49] B. Sanders, "Breast cancer with machine learning in MATLAB," *Texas Tech University Library*, págs. 1-400, dic. de 2019. dirección: https://ttu-ir.tdl.org/items/c71b307b-28ff-484c-af40-a79f37d52d57.

[50] A. Robert, *Machine Learning: The Complete Beginner's Guide to Learn and Effectively Understand Machine Learning Techniques (Intermediate, Advanced, To Expert Concepts)*. Amazon Digital Services LLC - KDP Print US, 2019, ISBN: 9781077867420. dirección: https://books.google.com.gt/books?id=YG0ixwEACAAJ.

[51] H. Greenspan, B. van Ginneken y R. M. Summers, "Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique," *IEEE Transactions on Medical Imaging*, vol. 35, n.º 5, págs. 1153-1159, 2016. DOI: 10.1109/TMI.2016.2553401.

[52] I. T. MathWorks, *Machine Learning con MATLAB*. dirección: https://la.mathworks.com/campaigns/offers/next/machine-learning-with-matlab.html.

[53] J. Suckling et al., "Mammographic Image Analysis Society (MIAS) database v1.21," 2015. DOI: 10.17863/CAM.105113. dirección: https://www.repository.cam.ac.uk/handle/1810/250394.

[54] T. M. Inc, *Filtrado de mediana de 2D*. dirección: https://la.mathworks.com/help/images/ref/medfilt2.html?searchHighlight=medfilt2&s_tid=srchtitle_support_results_1_medfilt2.

[55] T. M. Inc, *Binarizar una imagen 2D en escala de grises o un volumen 3D por medio del método del valor umbral*. dirección: https://la.mathworks.com/help/images/ref/imbinarize.html?s_tid=doc_ta.

[56] T. M. Inc, *Cerrar morfológicamente imágenes*. dirección: https://la.mathworks.com/help/images/ref/imclose.html?searchHighlight=imclose&s_tid=srchtitle_support_results_1_imclose.

[57] T. M. Inc, *Mejorar el contraste mediante la ecualización de histogramas*. dirección: https://la.mathworks.com/help/images/ref/histeq.html?s_tid=doc_ta.

[58] T. M. Inc, *Filtrado top-hat*. dirección: https://la.mathworks.com/help/images/ref/imtophat.html.

[59] I. T. MathWorks, *imshowpair*. dirección: https://la.mathworks.com/help/images/ref/imshowpair.html.

[60] W. Rahmaniar y W.-J. Wang, "Real-Time Automated Segmentation and Classification of Calcaneal Fractures in CT Images," *Applied Sciences*, vol. 9, n.º 15, 2019, ISSN: 2076-3417. DOI: 10.3390/app9153011. dirección: https://www.mdpi.com/2076-3417/9/15/3011.

[61] I. T. MathWorks, *Confusionchart*. dirección: https://la.mathworks.com/help/stats/confusionchart.html.

[62] D. A. Ragab, M. Sharkas, S. Marshall y J. Ren, "Breast cancer detection using deep convolutional neural networks and support vector machines," *PeerJ*, vol. 7, e6201, ene. de 2019, ISSN: 2167-8359. DOI: 10.7717/peerj.6201.

[63] P. I. R. Yoganapriya, "Pre-Processing Techniques for Digital Mammograms," *International Journal of Science and Research (IJSR)*, vol. 12, págs. 647-651, 2 feb. de 2023, ISSN: 23197064. DOI: 10.21275/SR23206162028.

[64] K. Sheba y S. G. Raj, "An approach for automatic lesion detection in mammograms," *Cogent Engineering*, vol. 5, pág. 1444320, 1 ene. de 2018, ISSN: 2331-1916. DOI: 10.1080/23311916.2018.1444320.

[65] by Khairul Nisak Bt Md Hasan y P. D. Ridzuan, *Detection of Microcalcification Using Mammograms*, 2004.

[66] T. M. Inc, *imclearborder*. dirección: `https://la.mathworks.com/help/images/ref/imclearborder.html`.

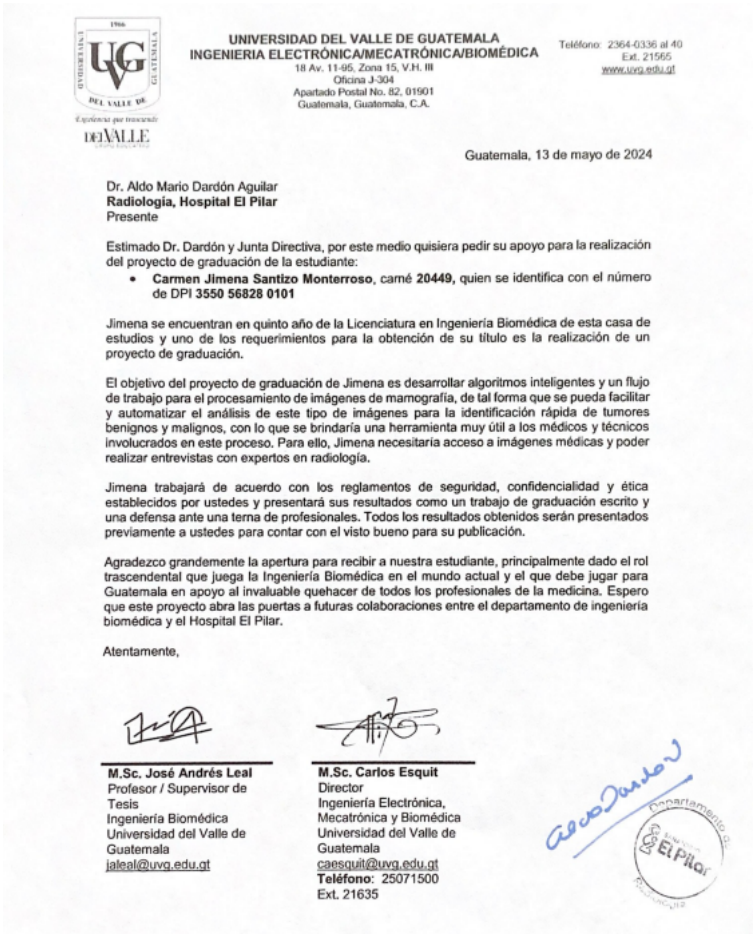
[67] T. M. Inc, *Procesamiento basado en ROI*. dirección: `https://la.mathworks.com/help/images/roi-based-processing.html`.

[68] T. M. Inc, *histcounts*. dirección: `https://la.mathworks.com/help/matlab/ref/double.histcounts.html`.

CAPÍTULO 13

Anexos

Figura 28. *Carta alianza UVG y Hospital El Pilar*



UNIVERSIDAD DEL VALLE DE GUATEMALA
INGENIERIA ELECTRÓNICA/MECATRÓNICA/BIOMÉDICA
18 Av. 11-95, Zona 15, V.H. III
Oficina J-304
Apartado Postal No. 82, 01901
Guatemala, Guatemala, C.A.

Teléfono: 2364-0336 al 40
Ext. 21565
www.uvg.edu.gt

Guatemala, 13 de mayo de 2024

Dr. Aldo Mario Dardón Aguilar
Radiología, Hospital El Pilar
Presente

Estimado Dr. Dardón y Junta Directiva, por este medio quisiera pedir su apoyo para la realización del proyecto de graduación de la estudiante:

- **Carmen Jimena Santizo Monterroso**, camé **20449,** quien se identifica con el número de DPI **3550 56828 0101**

Jimena se encuentran en quinto año de la Licenciatura en Ingeniería Biomédica de esta casa de estudios y uno de los requerimientos para la obtención de su título es la realización de un proyecto de graduación.

El objetivo del proyecto de graduación de Jimena es desarrollar algoritmos inteligentes y un flujo de trabajo para el procesamiento de imágenes de mamografía, de tal forma que se pueda facilitar y automatizar el análisis de este tipo de imágenes para la identificación rápida de tumores benignos y malignos, con lo que se brindaría una herramienta muy útil a los médicos y técnicos involucrados en este proceso. Para ello, Jimena necesitaría acceso a imágenes médicas y poder realizar entrevistas con expertos en radiología.

Jimena trabajará de acuerdo con los reglamentos de seguridad, confidencialidad y ética establecidos por ustedes y presentará sus resultados como un trabajo de graduación escrito y una defensa ante una terna de profesionales. Todos los resultados obtenidos serán presentados previamente a ustedes para contar con el visto bueno para su publicación.

Agradezco grandemente la apertura para recibir a nuestra estudiante, principalmente dado el rol trascendental que juega la Ingeniería Biomédica en el mundo actual y el que debe jugar para Guatemala en apoyo al invaluable quehacer de todos los profesionales de la medicina. Espero que este proyecto abra las puertas a futuras colaboraciones entre el departamento de ingeniería biomédica y el Hospital El Pilar.

Atentamente,

M.Sc. José Andrés Leal
Profesor / Supervisor de Tesis
Ingeniería Biomédica
Universidad del Valle de Guatemala
jaleal@uvg.edu.gt

M.Sc. Carlos Esquit
Director
Ingeniería Electrónica, Mecatrónica y Biomédica
Universidad del Valle de Guatemala
caesquit@uvg.edu.gt
Teléfono: 25071500
Ext. 21635

Departamento de El Pilar

Figura 29. *Parámetros de las imágenes mamográficas del dataset*

1st column: MIAS database reference number.

2nd column: Character of background tissue:
- F - Fatty
- G - Fatty-glandular
- D - Dense-glandular

3rd column: Class of abnormality present:
- CALC - Calcification
- CIRC - Well-defined/circumscribed masses
- SPIC - Spiculated masses
- MISC - Other, ill-defined masses
- ARCH - Architectural distortion
- ASYM - Asymmetry
- NORM - Normal

4thcolumn: Severity of abnormality;
- B - Benign
- M - Malignant

5 & 6th columns: x,y image-coordinates of centre of abnormality.

7th column: Approximate radius (in pixels) of a circle enclosing the abnormality.

Cuadro 1. *Datos asociados al dataset mamográfico, mamografías 01-25*

mdb001lm	CIRC	B	1815	1116	790	G
mdb002rl	CIRC	B	3091	1262	277	G
mdb003ll	NORM					D
mdb004rl	NORM					D
mdb005ll	CIRC	B	647	1163	122	F
mdb006rl	NORM					F
mdb007ll	NORM					G
mdb008rl	NORM					G
mdb009ll	NORM					F
mdb010rm	CIRC	B	2509	975	135	F
mdb011ll	NORM					F
mdb012rl	CIRC	B	2378	1467	162	F
mdb013ll	MISC	B	1574	1923	127	G
mdb014rl	NORM					G
mdb015lm	CIRC	B	3571	1359	275	G
mdb016rm	NORM					G
mdb017ls	CIRC	B	2407	943	192	G
mdb018rs	NORM					G
mdb019ll	CIRC	B	2021	1864	197	G
mdb020rl	NORM					G
mdb021ll	CIRC	B	612	1224	197	G
mdb022rm	NORM					G
mdb023ll	CIRC	M	2837	1405	117	G
mdb024rl	NORM					G
mdb025ll	CIRC	B	1886	1948	318	F

Cuadro 2. *Datos asociados al dataset mamográfico, mamografías 26-75*

mdb026rl	NORM					F
mdb027ll	NORM					F
mdb028rl	CIRC	M	2953	1999	224	F
mdb029ll	NORM					G
mdb030rm	MISC	B	1505	1785	174	G
mdb031ll	NORM					G
mdb032rl	MISC	B	1243	1798	267	G
mdb033ls	NORM					D
mdb034rs	NORM					D
mdb035ls	NORM					D
mdb036rs	NORM					D
mdb037ls	NORM					D
mdb038rs	NORM					D
mdb039ls	NORM					D
mdb040rs	NORM					D
mdb041ll	NORM					G
mdb042rl	NORM					G
mdb043ls	NORM					G
mdb044rs	NORM					G
mdb045lm	NORM					G
mdb046rm	NORM					G
mdb047lm	NORM					G
mdb048rm	NORM					G
mdb049ll	NORM					G
mdb050rl	NORM					G
mdb051ll	NORM					G
mdb052rm	NORM					G
mdb053ls	NORM					D
mdb054rs	NORM					D
mdb055lm	NORM					G
mdb056rm	NORM					G
mdb057ll	NORM					D
mdb058rl	MISC	M	2774	2079	110	D
mdb059ls	CIRC	B				F
mdb060rs	NORM					F
mdb061ls	NORM					D
mdb062rs	NORM					D
mdb063lm	MISC	B	1967	1163	133	D
mdb064rm	NORM					D
mdb065lm	NORM					D
mdb066rm	NORM					D
mdb067ll	NORM					D
mdb068rl	NORM					D
mdb069ll	CIRC	B	1739	1101	177	F
mdb070rl	NORM					F
mdb071lm	NORM					G
mdb072rm	ASYM	M	2140	2011	115	G
mdb073ls	NORM					G
mdb074rs	NORM					G
mdb075lm	ASYM	M	2982	850	92	F

Cuadro 3. *Datos asociados al dataset mamográfico, mamografías 76-125*

mdb076rm	NORM					F
mdb077ll	NORM					F
mdb078rl	NORM					F
mdb079lm	NORM					F
mdb080rm	CIRC	B	3615	1344	81	F
mdb081ll	ASYM	B	2007	1220	525	G
mdb082rl	NORM					G
mdb083ll	ASYM	B	891	1428	152	G
mdb084rl	NORM					G
mdb085lm	NORM					G
mdb086rm	NORM					G
mdb087lm	NORM					F
mdb088rm	NORM					F
mdb089lm	NORM					G
mdb090rm	ASYM	M	2021	1035	198	G
mdb091lm	CIRC	B	2090	1696	82	F
mdb092rm	ASYM	M	1562	1382	175	F
mdb093lm	NORM					G
mdb094rm	NORM					G
mdb095ll	ASYM	M	2181	1118	116	F
mdb096rl	NORM					F
mdb097ll	ASYM	B	1302	1702	137	F
mdb098rl	NORM					F
mdb099lm	ASYM	B	1473	1834	93	D
mdb100rm	NORM					D
mdb101lm	NORM					D
mdb102rm	ASYM	M	2369	1412	152	D
mdb103lm	NORM					D
mdb104rm	ASYM	B	2751	1645	203	D
mdb105ll	ASYM	M	1229	1318	392	D
mdb106rl	NORM					D
mdb107ll	ASYM	B	2597	1653	446	D
mdb108rl	NORM					D
mdb109ll	NORM					D
mdb110rl	ASYM	M	2502	2590	205	D
mdb111ll	ASYM	M	2414	1275	428	D
mdb112rl	NORM					D
mdb113ls	NORM					G
mdb114rs	NORM					G
mdb115ll	ARCH	M	2240	1096	468	G
mdb116rl	NORM					G
mdb117ll	ARCH	M	2417	1175	337	G
mdb118rl	NORM					G
mdb119ll	NORM					G
mdb120rl	ARCH	M	3162	1659	319	G
mdb121ll	ARCH	B	1849	1221	348	G
mdb122rl	NORM					G
mdb123lm	NORM					G
mdb124rm	ARCH	M	1729	1609	135	G
mdb125ll	ARCH	M	2322	2054	242	D

Cuadro 4. *Datos asociados al dataset mamográfico, mamografías 126-175*

mdb126rl	ARCH	B	2015	2585	93	D
mdb127lm	ARCH	B	2317	1069	194	G
mdb128rm	NORM					G
mdb129ll	NORM					D
mdb130rl	ARCH	M	2002	2469	112	D
mdb131lx	NORM					F
mdb132rx	CIRC	B	1499	3043	211	F
mdb133lx	NORM					F
mdb134rx	MISC	M	1736	2173	199	F
mdb135lx	NORM					F
mdb136rx	NORM					F
mdb137ll	NORM					D
mdb138rl	NORM					D
mdb139lx	NORM					F
mdb140rx	NORM					F
mdb141lx	CIRC	M	3591	1832	117	F
mdb142rx	CIRC	B	2104	2662	104	F
mdb143lx	NORM					F
mdb144rx	MISC	B	674	3117	119	F
mdb145lx	SPIC	B	2726	2631	197	D
mdb146rx	NORM					D
mdb147lx	NORM					F
mdb148rx	SPIC	M	2220	2745	699	F
mdb149lx	NORM					F
mdb150rx	ARCH	B	2005	2647	249	F
mdb151lx	NORM					F
mdb152rx	ARCH	B	2704	1349	195	F
mdb153lx	NORM					F
mdb154rx	NORM					F
mdb155ll	ARCH	M	2032	1046	380	F
mdb156rl	NORM					F
mdb157lm	NORM					F
mdb158rm	ARCH	M	1951	915	353	F
mdb159ll	NORM					F
mdb160rl	ARCH	B	2133	1206	245	F
mdb161lm	NORM					D
mdb162rm	NORM					D
mdb163ll	ARCH	B	1574	817	202	D
mdb164rl	NORM					D
mdb165ls	ARCH	B	2073	903	168	D
mdb166rs	NORM					D
mdb167ll	ARCH	B	2740	1550	141	F
mdb168rl	NORM					F
mdb169lm	NORM					D
mdb170rm	ARCH	M	2288	1118	331	D
mdb171ll	ARCH	M	2622	1102	248	D
mdb172rl	NORM					D
mdb173ll	NORM					F
mdb174rl	NORM					F
mdb175lm	SPIC	B	2795	1344	132	G

Cuadro 5. *Datos asociados al dataset mamográfico, mamografías 176-225*

mdb176rm	NORM					G
mdb177ls	NORM					G
mdb178rs	SPIC	M	1810	880	280	G
mdb179ls	SPIC	M	2168	1152	268	D
mdb180rs	NORM					D
mdb181lm	SPIC	M	1563	1052	217	G
mdb182rm	NORM					G
mdb183ll	NORM					F
mdb184rl	SPIC	M	1712	1943	458	F
mdb185ls	NORM					G
mdb186rs	SPIC	M	2114	1237	191	G
mdb187lm	NORM					G
mdb188rm	SPIC	B	1741	1448	247	G
mdb189ll	NORM					G
mdb190rl	SPIC	B	1724	1302	127	G
mdb191ls	SPIC	B	2177	1128	165	G
mdb192rs	NORM					G
mdb193ll	SPIC	B	2364	850	528	D
mdb194rl	NORM					D
mdb195ll	SPIC	B	631	2155	107	F
mdb196rl	NORM					F
mdb197lm	NORM					D
mdb198rm	SPIC	B	1761	800	373	D
mdb199lm	SPIC	B	820	1543	125	D
mdb200rm	NORM					D
mdb201ll	NORM					D
mdb202rl	SPIC	M	1122	1123	149	D
mdb203ll	NORM					F
mdb204rl	SPIC	B	2614	2005	84	F
mdb205ll	NORM					F
mdb206rl	SPIC	M	3410	1876	71	F
mdb207lm	SPIC	B	2370	1262	76	D
mdb208rm	NORM					D
mdb209ll	CALC	M	2126	1842	348	G
mdb210rl	NORM					G
mdb211lm	CALC	M	1423	1698	53	G
mdb212rm	CALC	B				G
mdb213ls	CALC	M	2193	940	183	G
mdb214rs	CALC	B				G
mdb215ll	NORM					D
mdb216rl	CALC	M				D
mdb217ll	NORM					G
mdb218rl	CALC	B	1694	1275	35	G
mdb219ll	CALC	B	3136	1439	119	G
mdb220rl	NORM					G
mdb221lm	NORM					D
mdb222rm	CALC	B	2502	1482	70	D
mdb223ls	CALC	B	2043	846	116	D
mdb224rs	NORM					D
mdb225lm	NORM					D

Cuadro 6. *Datos asociados al dataset mamográfico, mamografías 226-275*

mdb226rm	CALC	B	1770	1927	31	D
mdb227lm	CALC	B	1981	993	36	G
mdb228rm	NORM					G
mdb229ll	NORM					F
mdb230rl	NORM					F
mdb231ll	CALC	M	2265	1665	179	F
mdb232rl	NORM					F
mdb233lm	CALC	M				G
mdb234rm	NORM					G
mdb235ll	NORM					D
mdb236rl	CALC	B	912	2247	58	D
mdb237lm	NORM					F
mdb238rm	CALC	M	1998	986	70	F
mdb239ll	CALC	M	3133	1833	160	D
mdb240rl	CALC	B	1752	776	95	D
mdb241ls	CALC	M	2827	565	155	D
mdb242rs	NORM					D
mdb243lm	NORM					D
mdb244rm	CIRC	B	1940	1209	209	D
mdb245ls	CALC	M				F
mdb246rs	NORM					F
mdb247ll	NORM					F
mdb248rl	CALC	B	1805	1836	42	F
mdb249lm	CALC	M	2146	1154	194	D
mdb250rm	NORM					D
mdb251lm	NORM					F
mdb252rm	CALC	B	2743	1318	94	F
mdb253ll	CALC	M	2368	2185	112	D
mdb254rl	NORM					D
mdb255ll	NORM					F
mdb256rl	CALC	M	2272	1750	149	F
mdb257ll	NORM					D
mdb258rl	NORM					D
mdb259ll	NORM					D
mdb260rl	NORM					D
mdb261ls	NORM					D
mdb262rs	NORM					D
mdb263lm	NORM					G
mdb264rm	MISC	M	2487	691	147	G
mdb265lm	MISC	M	2104	1351	242	G
mdb266rm	NORM					G
mdb267ll	MISC	M	2036	2427	227	F
mdb268rl	NORM					F
mdb269lm	NORM					G
mdb270rm	CIRC	M	430	1649	291	G
mdb271ll	MISC	M	1193	2391	274	F
mdb272rl	NORM					F
mdb273ll	NORM					F
mdb274rx	MISC	M	2630	3542	495	F
mdb275ll	NORM					G

Cuadro 7. *Datos asociados al dataset mamográfico, mamografías 276-322*

mdb276rl	NORM					G
mdb277lm	NORM					G
mdb278rm	NORM					G
mdb279ll	NORM					G
mdb280rx	NORM					G
mdb281lm	NORM					D
mdb282rm	NORM					D
mdb283lm	NORM					D
mdb284rm	NORM					D
mdb285lm	NORM					D
mdb286rm	NORM					D
mdb287ls	NORM					D
mdb288rs	NORM					D
mdb289ls	NORM					D
mdb290rs	CIRC	B	2799	1502	181	D
mdb291ll	NORM					G
mdb292rl	NORM					G
mdb293ll	NORM					F
mdb294rl	NORM					F
mdb295ll	NORM					D
mdb296rl	NORM					D
mdb297ll	NORM					F
mdb298rl	NORM					F
mdb299ll	NORM					F
mdb300rl	NORM					F
mdb301lm	NORM					F
mdb302rm	NORM					F
mdb303lm	NORM					F
mdb304rm	NORM					F
mdb305lm	NORM					F
mdb306rm	NORM					F
mdb307ll	NORM					F
mdb308rl	NORM					F
mdb309ll	NORM					F
mdb310rl	NORM					F
mdb311ll	NORM					F
mdb312rl	MISC	B	3158	2389	81	F
mdb313ll	NORM					F
mdb314rl	MISC	B	3447	1277	158	F
mdb315ll	CIRC	B	1900	1317	372	D
mdb316rl	NORM					D
mdb317ls	NORM					D
mdb318rs	NORM					D
mdb319ll	NORM					D
mdb320rl	NORM					D
mdb321lm	NORM					D
mdb322rm	NORM					D

Figura 30. *Etapa de posprocesamiento y segmentación. Visualización dos canales (verde y magenta) de la mamografía. Extracción de características*

```
%% FILTRADO DE POST PROCESAMIENTO
se2 = strel('disk', 20);
tophatFiltered = imtophat(Ieq, se2);
contrastAdjusted = imadjust(tophatFiltered);
level2 = graythresh(contrastAdjusted);
bw2 = imbinarize(contrastAdjusted, level2);
imshowpair(I_gray, contrastAdjusted, 'montage');
imshowpair(I_original, contrastAdjusted, 'montage');

noBorder = imclearborder(Ieq);
pectoralOnly = xor(noBorder, Ieq);
imshowpair(I_gray, pectoralOnly, 'montage');
imshowpair(I_original, pectoralOnly, 'montage');

%% VISUALIZACIÓN DE TUMOR (en caso exista)
imshowpair(pectoralOnly, ~contrastAdjusted, 'falsecolor', ...
    'ColorChannels', 'green-magenta', ...
    'Interpolation', 'nearest');
title(tipo);
axis off;

%% EXTRAER CARACTERÍSTICAS
% Convertir a escala de grises si es necesario
if size(pectoralOnly, 3) > 1
    pectoralOnly_gray = rgb2gray(pectoralOnly);
else
    pectoralOnly_gray = pectoralOnly; % Ya es de un solo canal
end

if size(contrastAdjusted, 3) > 1
    contrastAdjusted_gray = rgb2gray(contrastAdjusted);
else
    contrastAdjusted_gray = contrastAdjusted; % Ya es de un solo canal
end

% Extraer características del canal pectoralOnly
lbp_pectoral = extractLBPFeatures(pectoralOnly_gray);
hist_pectoral = histcounts(pectoralOnly_gray, 256); % Histograma del canal pectoral

% Extraer características del canal contrastAdjusted
lbp_contrast = extractLBPFeatures(contrastAdjusted_gray);
hist_contrast = histcounts(contrastAdjusted_gray, 256);

% Binarización
pectoralOnly_gray_uint8 = uint8(pectoralOnly_gray * 255);
threshold = graythresh(pectoralOnly_gray_uint8);
binaryImage = imbinarize(pectoralOnly_gray_uint8, threshold);
```

Figura 31. *Clasificación. Procesamiento de etiquetas asociadas a las características extraídas. División del conjunto de prueba y entrenamiento. Entrenamiento del modelo SVM y obtención de matriz de confusión*

```
    % Extraer características de forma
    stats = regionprops(binaryImage, 'Area', 'Perimeter', 'Eccentricity', 'Solidity');
    if ~isempty(stats)
        area = stats.Area;
        perimeter = stats.Perimeter;
        eccentricity = stats.Eccentricity;
        solidity = stats.Solidity;
    else
        area = 0;
        perimeter = 0;
        eccentricity = 0;
        solidity = 0;
    end
    % Características de intensidad
    mean_intensity = mean(pectoralOnly_gray(:));
    std_intensity = std(double(pectoralOnly_gray(:)));
    % Concatenar características
    features_shape = [lbp_pectoral, hist_pectoral, lbp_contrast, hist_contrast, ...
    area, perimeter, eccentricity, solidity, mean_intensity, std_intensity];
    % Agregar a la matriz de características
    extractedFeatures = [extractedFeatures; features_shape];
    end
% Verificar que 'extractedFeatures' no esté vacío
if isempty(extractedFeatures)
    error('No se han extraído características, verifica el proceso de extracción.');
end
% Procesar las etiquetas
labels = grp2idx(data.Gravedad_de_anomalia); % convierte tipos a índices
cv = cvpartition(labels, 'HoldOut', 0.2); % 80% entrenamiento, 20% prueba
idx = cv.test;

% Dividir en conjuntos de entrenamiento y prueba
dataTrain = extractedFeatures(~idx, :);
dataTest = extractedFeatures(idx, :);
labelTrain = labels(~idx);
labelTest = labels(idx);

% Contar la cantidad de muestras por clase
disp('Distribución de clases en el conjunto de entrenamiento:');
disp(countcats(categorical(labelTrain)));
disp('Distribución de clases en el conjunto de prueba:');
disp(countcats(categorical(labelTest)));

% Entrenar un modelo SVM
mdl = fitcsvm(dataTrain, labelTrain);
predictions = predict(mdl, dataTest);
accuracy = sum(predictions == labelTest) / length(labelTest);
fprintf('Precisión: %.2f%%\n', accuracy * 100);

% Visualizar la matriz de confusión
confMat = confusionmat(labelTest, predictions);
confusion_matrix = confusionchart(labelTest, predictions);
title('Matriz de Confusión');
```