Trabajo de Investigación

DETECCIÓN AUTOMÁTICA DE MICROCALCIFICACIONES EN MAMOGRAFÍAS DIGITALES

Felipe Palta

Documento presentado a
Programa de Posgrado en Ingeniería Eléctrica y Electrónica
como requisito para optar por el título de
Magíster en Ingeniería Énfasis en Electrónica

Directores: Humberto Loaiza Correa, Ph.D.
Sandra Esperanza Nope Rodríguez, Ph.D.

Escuela de Ingeniería Eléctrica y Electrónica
Facultad de Ingeniería

Octubre 2022

Resumen

En este documento se presenta el desarrollo de un sistema de visión artificial para la detección automática de microcalcificaciones en mamografías digitales utilizando máquinas de aprendizaje superficial. Se plantea y ejecuta una metodología partiendo de las etapas básicas de la construcción de un sistema de visión tradicional, empezando por la selección de la base datos MIAS.

Luego, se emplea una etapa de preprocesamiento y segmentación del área de interés para así extraer, a partir de ventanas previamente recortadas, las características más apropiadas para la discriminación de las clases y seleccionar las mejores para finalmente implementar una serie de clasificadores: RNA, SVM, K-NN y NB basado en un esquema de pruebas, estableciendo un umbral mínimo de 95 % de exactitud para las primeras tres y un mínimo de 80 % de exactitud para NB, para así seleccionar los mejores modelos de acuerdo a un análisis de varianza ANOVA.

Posteriormente, se evalúa el desempeño de esta solución, utilizando métricas apropiadas para *datasets* desbalanceados como lo son la precisión y la sensibilidad, ponderando esta última con el fin de contrarrestar el efecto del desbalance dentro del *ground truth* proporcionado por los expertos, obteniendo un 99.8 % de precisión y un 77 % de sensibilidad para una imagen mamográfica completa.

Palabras clave− aprendizaje superficial, cáncer de mama, mamografías digitales, visión artificial, microcalcificaciones.

Abstract

This paper presents the development of a computer vision system for the automatic detection of microcalcifications in digital mammograms using shallow learning machines. A methodology is proposed and executed starting from the basic stages of the construction of a traditional vision system, beginning with the selection of the MIAS database.

Then, a stage of preprocessing and segmentation of the area of interest is implemented in order to extract, from previously cut windows, the most appropriate features for class discrimination and select the best ones to finally implement a series of classifiers: ANN, SVM, K-NN and NB based on a testing scheme, using a minimum threshold of 95 % accuracy for the first three and a minimum of 80 % accuracy for NB, in order to select the best models according to an ANOVA variance analysis.

Subsequently, the performance of this solution was evaluated, using appropriate metrics for unbalanced datasets such as accuracy and sensitivity, weighting the latter in order to counterbalance the effect of the unbalance within the ground truth provided by the experts, obtaining 99.8 % accuracy and 77 % sensitivity for a complete mammographic image.

Keywords– machine learning, breast cancer, digital mammogram, computer vision, microcalcificacions

Agradecimientos

A mis directores los profesores Humberto Loaiza y Sandra Nope, quienes me orientaron en este trabajo y contribuyeron para que éste terminara satisfactoriamente. Gracias también por las enseñanzas y la confianza depositada.

A mis amigos y compañeros del Laboratorio de Visión Artificial del grupo de Investigación Percepción y Sistemas Inteligentes (PSI) de la Universidad del Valle, quienes aportaron con sus conocimientos y recomendaciones a lo largo del desarrollo de esta investigación.

A mi familia por su amor y apoyo constante a lo largo de mi vida académica, pero en especial a mi señora madre y dos tías, núcleo familiar principal, quienes siempre me han apoyado incondicionalmente y desean verme un profesional ejemplar llegando a lo más alto.

A la Universidad del Valle y su excelente planta docente, la cual me permitió conocer e involucrarme en la investigación científica de manera apasionante.

Este trabajo de investigación se desarrolló con recursos aportados por:

- Universidad del Valle mediante el programa de Asistencia de Docencia 2019-2020.

Glosario

ANOVA Análisis de la varianza. 54, 94–96

AUC Area bajo la curva. 9

BCDR Repositorio Breast cancer digital. 8, 13, 20

BI-RADS Breast imaging report and database system. 2

BPCS Segmentación compleja de planos de bits. 33, 77

CAD Detección asistida por computadora. 5

CCL Etiquetado de componentes conectados. 11, 20, 27, 74, 75, 99

CDI Carcinoma ductal infiltrante. 1

CDIS Carcinoma ductal in situ. 1

CIFE Selección de características basadas en la extracción condicional de características infomax. 10, 20

CM Cáncer de mama. 1–3, 5, 6, 8, 9, 15

CMIM Selección de características basadas en maximización de la información mutua condicional. 10, 20

CNN Redes neuronales convolucionales. 6, 11, 20

DDSM Base de datos Digital Database for Screening Mammography. 20

DISR Selección de características basadas en la relevancia simétrica de doble entrada. 10, 20

DRBM Máquina de Boltzmann restringida discriminatoria. 11, 20

DWT Transformada discreta de Wavelet. 10, 20

EF Fracción de superposición adicional. 12

EJECALS Base de datos EJECALS. 20

FFDM Mamografía digital de campo completo. 8, 13, 14, 19, 20

FSCR Selección de características basadas en la proporción de Fisher. 10, 20

GINI Selección de características basadas en el índice de Gini. 10, 20

GLCM Matriz de características de co-ocurrencia de nivel de grises. 9, 10, 19, 20, 41

GLRLM Matriz de características de longitudes de nivel de grises. 9, 20, 46, 86

GRA Características basadas en el gradiente. 9, 20

HT Transformada de Hough. 8

ICAP Selección de características basadas en límite de interacción. 10, 20

JMI Selección de características basadas en la técnica de información mutua-conjunta. 10, 20

K-NN K-vecinos cercanos. 55, 94, 95, 133

LBP Patrones binarios locales. 7, 10, 20, 44

LBPU Patrones binarios locales uniformes. 86

LDA Análisis discriminante lineal. 10, 20

MCs Microcalcificaciones. 1–3, 5–7, 10–15, 19, 55, 57–59, 77, 82–85, 87, 107, 108, 110–112, 114, 116, 118, 119, 121, 123, 133, 136

MIAS Base de datos Mammographic Image Analysis Society. 9, 11, 12, 20, 56, 57, 65, 66, 98, 136

MIM Selección de características basadas maximización de la información mútua. 10, 20

miniMIAS Base de datos reducida Mammographic Image Analysis Society. 9–12, 20

MLPC Red neuronal perceptrón multicapa. 10, 20

MRMR Selección de características basadas en mínima redundancia y máxima redundancia. 10, 20, 47, 48, 87, 88

NaN Not a Number. 86, 87

NB Modelo bayesiano. 10, 20, 55, 94, 95

NCA Neighbourhood components analysis. 89

NMS Técnica de supresión no máxima. 11, 20

REFL Selección de características basadas en la técnica Relief. 10, 20

RF Modelo de bosque aleatorio. 10, 11, 20

RNA Redes neuronales artificiales. 6, 51, 55, 94, 133

ROI Región de interés. 7–11, 15, 20, 31, 34

SFA Características basadas en el análisis espacial y frecuencial. 9, 20

SFM Mamografía convencional de película digitalizada. 8, 13, 14, 19, 20, 65

SFTA Segmentación con análisis de textura fractal. 7, 20, 47, 86

SI Indice de similaridad. 12

SNUBH-MDB Base de datos mamográfica del hospital Bundang de la Universidad Nacional de Seúl. 11, 12, 20

STA Características estadísticas. 9, 10, 19, 20

SVM Máquinas de vectores de soporte. 6, 7, 10, 20, 55, 94, 95, 133

SWM Método de deslizamiento de ventana. 11, 13, 15, 20

TTBD Descomposición binaria de dos umbrales. 7

Tabla de contenido

Pag.

ÍNDICE DE FIGURAS

FIGURA Pag.

ÍNDICE DE TABLAS

TABLA Pag.

1. Introducción

El carcinoma ductal in situ (CDIS) y el carcinoma ductal infiltrante (CDI) son patologías de alto riesgo invasivo en las mamas y son el tipo de cáncer de mama (CM) más común en mujeres alrededor del mundo. Estos se generan cuando existe una proliferación excesiva de células sobre los conductos galactóforos, por donde circula la leche materna producida por las glándulas mamarias [American Cancer Society, 2017].

Uno de los primeros métodos de identificación y posible diagnóstico del CM es la mamografía, prueba realizada anualmente a mujeres que llegan a la etapa postmenopáusica. La mamografía también puede ser ordenada a mujeres que no hayan llegado a la menopausia cuando se presenta sintomatología sospechosa en las mamas o simplemente cuando se posee un antecedente familiar relacionado con el CM. Indicadores de primer orden en mamografías que representan signos primarios de un posible tumor maligno, son el hallazgo de condensaciones espiculadas de bordes mal definidos, distorsiones de la arquitectura, opacidades diferenciadas o asimetrías del parénquima.

Otros indicadores, llamados de segundo orden, son la presencia de microcalcificaciones (MCs), que representan un dato radiológico en la detección del CM en estadios iniciales. Muchos de los autores coinciden en que la presencia de MCs puede indicar un CM oculto, asociando que cuando éstas aparecen de manera coligada a una masa o distorsiones de la arquitectura glandular, es el mayor valor predictivo de malignidad, mientras que cuando aparecen de manera aislada o separada de otras lesiones no revisten un indicador predictivo de malignidad [A. Jáuregui et al., 1994].

Por tanto, el análisis de las MCs representa uno de los grandes retos para la radiología cuando del CM se trata, ya que son el signo más frecuente de cáncer en un estadio temprano y pueden verse incluso antes de que una masa sea palpable o detectada radiológicamente.

Las MC son visualmente reconocidas en mamografías como puntos blancos individuales o agrupados de un patrón específico y asociadas con grados de severidad

benigna o maligna mediante los descriptores BI-RADS (Breast Imaging Report and Database System) [Arancibia et al., 2016].

De acuerdo con la extensión del CM, se pueden presentar alguno de los siguientes estadios: 0, hay células anormales sin diseminación al tejido cercano; I, II, III, presencia de cáncer y, entre mayor sea el número asociado, es más grande el tamaño y la extensión del tumor; IV el cáncer se encuentra diseminado a partes distantes del cuerpo [BreastCancerOrganization, 2018].

Un médico radiólogo realiza su primer diagnóstico de CM utilizando principalmente mamografías, es decir, imágenes del seno en el espectro de rayos X. Estos exámenes de manera preventiva en Colombia, inician desde los 40 años hasta los 74 años, exceptuando casos con antecedentes clínicos o hereditarios relacionados con el CM, y se toman con repeticiones anuales, siguiendo las recomendaciones directas de las asociaciones médicas de los Estados Unidos, en especial el Grupo de Trabajo de Servicios Preventivos de Estados Unidos (USPSTF) [Academia Nacional de Medicina de Colombia, 2016]. Cuando se utiliza el diagnóstico mediante mamografías de forma exclusiva, se han reportado tasas de detección que fluctúan con la edad del paciente; así, en mujeres entre los 40 y 50 años se detecta entre el 75 y el 90 % de los cánceres de mama; mientras que en mujeres entre los 50 y 60 años la tasa aumenta entre el 94 y 97 % [K. Pesce et al., 2012]. La diferencia en la efectividad se debe fundamentalmente a la disminución de la densidad del tejido mamario en pacientes mayores a 60 años, lo que contribuye a que las MCs sobresalgan de manera más clara en la imagen mamográfica, por tanto, haciendo más fácil su detección visual del médico radiólogo y su asociación a patologías mamarias.

Una de las razones por las que la localización y caracterización del CM en mamografías es compleja, es debido a que las MCs asociadas al cáncer y a las comúnmente benignas, presentan pocas variaciones morfológicas (tamaño, contextura, agrupación y contraste), por lo que tienden a confundirse con el tejido mamario. De este modo, el éxito en la detección está fuertemente ligado a la experticia del médico y a factores psicofisiológicos (fatiga, estrés, problemas visuales, etc.), que se potencializan por el alto volumen de mamografías para ser analizadas frente a una baja disponibilidad de especialistas en los centros médicos. El estudio realizado por [Lee et al., 2017], evidencia que teniendo diez radiólogos de distintas instituciones

médicas examinando una base de datos de mamografías, existe una alta variabilidad entre éstos en la determinación del estadio BI-RADS, concluyendo así que para evaluar el valor predictivo de ciertas características mamográficas para la malignidad deben tomar en consideración la variabilidad interpretativa inherente de estos hallazgos.

En la actualidad, la implementación de técnicas de visión artificial para la automatización de los diferentes procesos realizados dentro de los centros médicos es una prioridad, pues conllevan a un aumento de la productividad, mejora en la calidad de los diagnósticos y es una forma de innovar el área médica [Esteva et al., 2019].

Una estrategia para volver más eficiente los diagnósticos en las imágenes radiológicas de mamografías, es utilizar sistemas de procesamiento digital de imágenes, detectar zonas con posibles MCs mediante sus características, y marcarlas para que el médico radiólogo las analice sin pasar por alto alguna de ellas, por tanto, cuente con una herramienta de apoyo que mejore la detección temprana del CM y a su vez; ayude a reducir los tiempos requeridos en el análisis de las imágenes mamográficas.

Con base en lo anterior, en este trabajo se planteó como pregunta de investigación: **¿Cómo detectar microcalcificaciones en mamografías digitales utilizando técnicas de visión artificial?**

Para resolver el problema propuesto se definió un objetivo general y cuatro específicos que se indican a continuación:

Objetivo general

Desarrollar un sistema de visión artificial para la detección de microcalcificaciones en mamografías.

Objetivos específicos

- Seleccionar las características discriminantes en la identificación de microcalcificaciones medibles a partir de mamografías digitales.

- Implementar un algoritmo para segmentar las regiones con posible existencia de microcalcificaciones en mamografías digitales.
- Desarrollar un algoritmo para clasificar e identificar microcalcificaciones en mamografías digitales.
- Determinar los alcances y limitaciones del sistema en la detección de microcalcificaciones.

2. Antecedentes

En este capítulo se explora el panorama investigativo a nivel nacional e internacional en la temática de detección automática de MCs en mamografías. Esta información se sintetizó a partir de los filtros de búsqueda que permitieron definir el impacto en la detección de MCs y la cantidad de publicaciones por año utilizando recursos web de publicaciones indexadas.

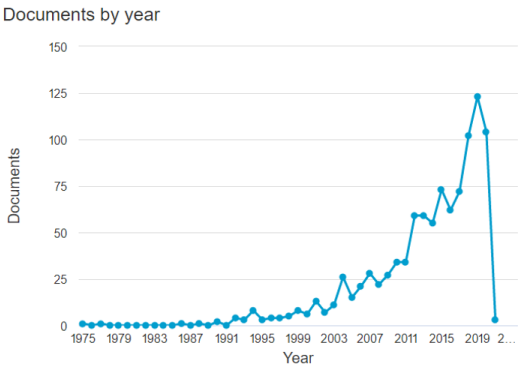
2.1 Panorama global

El CM es uno de los más comunes que afectan principalmente a mujeres alrededor del mundo, y el examen visual de la mamografía es uno de los métodos principales para ayudar en su detección. Sistemas de detección asistida por computadora (CAD), que se usan para detectar el CM en mamografías ayudan en la reducción de errores humanos. Los sistemas CAD comerciales son desarrollados e implementados exclusivamente en los mamografos, lo cual no permite el acceso al público en general.

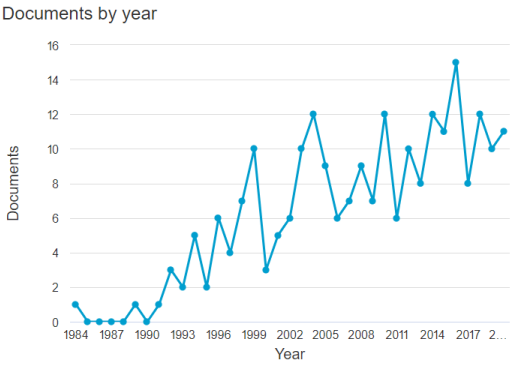
En las últimas dos décadas y a raíz del aumento de la capacidad de procesamiento ofrecido por los sistemas de cómputo modernos, la literatura científica presenta un incremento sustancial en publicaciones con desarrollos orientados a la detección del cáncer de mama haciendo uso de sistemas CAD. Por tanto, la detección de MCs en mamografías digitales de manera automática haciendo uso de sistemas de visión artificial, presenta también un grupo significativo de investigaciones y aportes en la comunidad científica.

Haciendo uso de la herramienta de análisis de datos de Scopus Elsevier, Las Figuras 2.1a y 2.1b muestran el incremento de publicaciones en **Detección Automática de Cáncer de Mama en Mamografías** y **Detección Automática de Microcalcificaciones en Mamografías** respectivamente a lo largo de las útlimas tres décadas alrededor del mundo.

En la actualidad, la detección Automática de MCs se realiza mediante el uso de



(a) Publicaciones de Detección Automática de Cáncer de Mama en Mamografías [Fuente de datos: Scopus].



(b) Publicaciones de Detección Automática de Microcalcificaciones en Mamografías [Fuente de datos: Scopus].

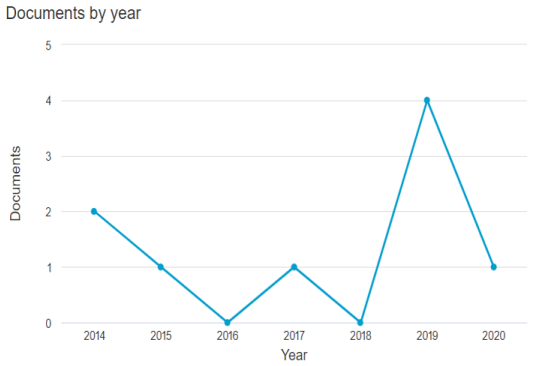
Figura 2.1: Publicaciones globales relacionadas con detección automática de cáncer de mama y microcalcificaciones.

diferentes metodologías basadas en técnicas de Visión por Computador y Aprendizaje Automático como: Preprocesamiento de imagenes, Segmentación, Transformaciones espaciales o frecuenciales y clasificación mediante máquinas de aprendizaje automático clásicas como Redes Neuronales Artificiales (RNA), Máquinas de Vectores de Soporte (SVM), entre otras, o de aprendizaje profundo como Redes Neuronales Convolucionales (CNN). En la lituratura científica también se observa el uso mixto de los múltiples tipos de técnicas mencionadas.

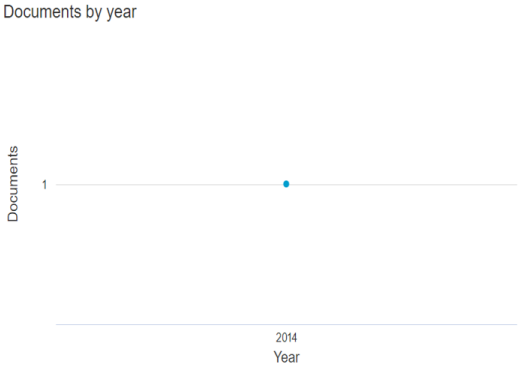
2.2 Antecedentes Nacionales

A nivel nacional el panorama investigativo es escaso con respecto a trabajos relacionados en la detección automática de MCs, por otro lado, hay mayor concentración de trabajos enfocados a la detección automática de las múltiples patologías del CM. En ambos casos, la información existente, aporta un contenido científico relevante en el área. Las Figuras 2.2a y 2.2b presentan las publicaciones realizadas en Colombia acerca de **Detección Automática de Cáncer de Mama en Mamografías** y **Detección Automática de Microcalcificaciones en Mamografías** respectivamente.

La técnica de reconocimiento automático de MCs reportada en [Vargas et al.,



(a) Publicaciones de Detección Automática de Cáncer de Mama en Mamografías en Colombia [Fuente de datos: Scopus].



(b) Publicaciones de Detección Automática de Microcalcificaciones en Mamografías en Colombia [Fuente de datos: Scopus].

Figura 2.2: Publicaciones relacionadas con detección automática de cáncer de mama y microcalcificaciones en Colombia.

2014] presenta un pre-procesamiento de la imagen mamográfica haciendo uso el operador morfológico de erosión con el objetivo de segmentar el contorno del seno, donde posteriormente realiza un filtrado de desenfoque para resaltar la región de interés (ROI), a la cual le realiza la extracción de características mediante la segmentación con análisis de textura fractal (SFTA) generando un conjunto de imágenes binarias mediante la descomposición binaria de dos umbrales (TTBD). El algoritmo SFTA genera un vector de características con el tamaño, nivel de grises y contorno de las imágenes binarias. Se entrena un modelo SVM 50 veces de manera aleatoria y se comparan los resultados de clasificación con las características extraidas del método patrones binarios locales (LBP) encontrando una exactitud mucho mayor en los resultados provenientes de las características SFTA que LBP en las tres clases: **Ben** - microcalcificación benigna, **Mal** - microcalcificación maligna y **NoC** - no microcalcificación; con las siguientes cuatro combinaciones evaluadas: 1. MCs benignas y malignas, 2. MCs benignas e imagen sin MCs, 3. MCs malignas e imagen sin MCs y 4. MCs benignas-malignas e imagen sin MCs. El trabajo realizado resalta que la técnica SFTA supera a LBP conluyendo que la primera cuantifica mejor la información de textura en una mamografía con presencia de MCs, sin embargo, se debe de tener en cuenta que la selección de los umbrales alto y bajo de TTBD se realizarón a modo prueba y error, lo cual implica que una selección incorrecta de éstos, puede afectar fuertemente el desempeño de los resultados. La base de datos

EJECALS, es local y proviene de la colaboración de la fundación Alejandro Londoño, en el sector del eje cafetero.

Por otro lado, en [Pertuz et al., 2019] se presenta un sistema estructurado automático para la evaluación del riesgo de CM usando la base de datos comúnmente utilizada *BREAST CANCER DIGITAL REPOSITORY - BCDR* ([University of Aveiro, 2008]) en múltiples investigaciones relacionadas con la detección automática del CM. El repositorio BCDR está compuesto principalmente por dos tipos de imagenes: 1. Mamografía convencional de película digitalizada (SFM) y 2. Mamografía digital de campo completo (FFDM). Aunque ambos métodos de captura de mamografías estan basados en el uso de rayos X para producir la imagen de la mama, la diferencia radica en que SFM es almacenada y visualizada en un pelicula (*papel térmico*), mientras que FFDM es almacenada y visualizada en un sistema de cómputo. El reto principal en la utilización de una mamografía SFM, es eliminar la presencia de ruido y artefactos sobre la imagen provenientes del proceso de digitalización. El repositorio BCDR posee información avalada por expertos, donde las imágenes están categorizadas según el nivel de riesgo: alto o bajo. Como es común en el desarrollo de sistemas de detección y evaluación automáticos basados en algoritmos de visión por computador, el autor propone un protocolo basado en cuatro etapas fundamentales: a. **segmentación del seno**, b. **detección de la ROI**, c. **extracción de características** y d. **clasificación (evaluación del riesgo de CM)**. La ***segmentación del seno-(a)*** a su vez comprende tres tareas principales: a1. detección del fondo, a2. detección del tejido pectoral y a3. detección del pezón. Para la detección del fondo (a1), se proponen un método estableciendo un umbral de entrada basado en la intensidad del histograma (usado en las imagenes FFDM) y otro método estadístico basado en las intensidades de cada pixel, los cuales son evaluados con el test Anderson-Darling para identificar la región del fondo y el seno (usado en las imagenes SFM). Para la detección del tejido pectoral (a2), se aplica un detector de borde a la imagen de entrada y mediante la transformada de Hough (HT) cada pixel del borde es representado en el espacio acumulador de Hough (usado en ambas imagenes FFDM y SFM). La detección del pezón (a3) se basa en el resultado de (a2), tomando el punto más alejado del trazo del pectoral encontrado. La ***detección de la ROI-(b)*** parte del resultado la *segmentación del seno-(a)* y tiene como objetivo detectar la región específica dentro del seno para realizar la posterior extracción de características. El autor propone utilizar cuatro tipos de ROI: b1. toda la región del

seno, sugerida como la más usada en [Nielsen et al., 2014], b2. región cuadrada seleccionada manualmente, b3. región retroareolar desarrollada por el autor en [Torres and Pertuz, 2016], el cual se basa en un sistema de coordenadas curvilíneas que se adapta automáticamente a la forma y tamaño del seno y b4. cuadrícula de múltiples regiones espaciadas dentro del seno, técnica presentada en [Zheng et al., 2015]. La ***extracción de características-(c)*** es llevada a cabo tratando de recopilar las cinco técnicas de extracción de características usualmente utilizadas en el campo de visión artificial: c1. características estadísticas (STA), debido a que generalmente las intensidades de los pixeles están relacionadas a la radiodensidad del tejido y son asociadas al riesgo de CM, son fáciles de computar y generan un buen desempeño en algunas imágenes médicas como las mamográficas; c2. características de textura de la imagen basadas en la matriz de co-ocurrencia del nivel de grises (GLCM), c3. características de textura basadas en la matriz de longitudes del nivel de grises (GLRLM), c4. características basadas en el gradiente (GRA) y c5. características basadas en el análisis espacial y frecuencial (SFA). Finalmente, la ***clasificación-(d)***, se realizó utilizando un clasificador basado en regresión logística con selección de características por pasos, donde se busca de manera exhaustiva todas las características disponibles y compara el rendimiento con la regresión. El resultado obtenido es un valor numérico $r \in [0, 1]$ donde 0 es el valor de menor riesgo y 1 el valor de mayor riesgo. Los resultados obtenidos, indican que el mejor desempeño en la clasificación se dió con ROI región cuadrada seleccionada manualmente (b2) y haciendo uso de todas las características en el proceso de adecuación del modelo y posterior validación, alcanzando un valor máximo de área bajo la curva (AUC) de 0.846.

Por otro lado, el trabajo presentado en [Rincón et al., 2018] propone un sistema para la clasificación de anormalidades en las regiones de una imagen mamográfica. La base de datos utilizada es *Mammographic Image Analysis Society - (MIAS* 50 μm *por pixel)* en su versión reducida (miniMIAS 200 μm por pixel). Nuevamente, el autor presenta un protocolo basado en las etapas fundamentales del agortimos para la clasificación basados en visión por computador. Las etapas son: a. **selección de la ROI**, b. **extracción de características**, c. **selección de las mejores características** y d. **clasificación (evaluación de anormalidad)**. La ***selección de la ROI-(a)*** para el proceso de posterior entrenamiento fué realizado de manera manual teniendo en cuenta las anotaciones realizadas por los expertos de la base de

datos miniMIAS y definiendo una ROI estandar de 128 x 128 *pixeles* generando tres clases: **benigna**, **maligna** y **tejido normal**. Se extrajeron en total 91 ROIs y se resalta que fué excluida la presencia de MCs. Con el objetivo de evaluar el efecto de múltiples descriptores de textura para representar el contenido de la ROI, la ***extracción de características-(b)*** se basó en usar la técnica de descomposición multiresolución usando la transformada discreta de Wavelet (DWT) y obteniendo las coeficientes: horizontal, vertical y diagonales de la ROI. A partir de los coeficientes, el autor propone la obtención de características usando tres técnicas: b1. trece características estadísticas de primer orden STA, b2. trece característiacs a partir del cómputo de la matriz de co-ocurrencia del nivle de grises GLCM a la distancia unidad y en sus ángulos principales: 0, 45, 90, y 125° y b3. patrones binarios locales LBP. Para la ***selección de las mejores características-(c)***, el autor evalúa los diez métodos más utilizados en el repositorio scikit-feature perteneciente a la librería Scikit-learn y presentados en [Jundong Li et al., 2017], comúnmente utilizados para la solución de problemas de aprendizaje de automático. Estos son: c1. proporción de Fisher (FSCR), c2. técnica Relief (REFL), c3. índice de Gini (GINI), c4. técnica de información mútua-conjunta (JMI), c5. extracción condicional de características infomax (CIFE), c6. reelevancia simétrica de doble entrada (DISR), c7. maximización de la información mútua (MIM), c8. maximización de la información mutua condicional (CMIM), c9. límite de interacción (ICAP) y c10. mínima redundancia y máxima redundancia (MRMR). Aplicando las diez técnicas a las característiacs, de manera reiterativa se pudo encontrar que las características GLCM y LBP generaban la mejor separabilidad entre clases para su posterior clasificación. En el proceso de ***clasificación (evaluación de anormalidad)-(d)***, el autor utiliza cinco modelos comúnmente usados para la clasificación de anormalidades en imagenes médicas como se menciona en [Erickson et al., 2017] y estos son: d1. Máquinas de vectores de soporte SVM, d2. red neuronal perceptrón multicapa (MLPC), d3. modelo de bosque aleatorio (RF), d4. modelo bayesiano (NB) basado en el clasificador ingenuo de Bayes y d5. clasificación mediante el uso de la técnica de análisis discriminante lineal (LDA). Finalmente la evaluación del desempeño se realizó mediante la estrategia de validación cruzada, donde la mejor precisión encontrada fue utilizando SVM y MLPC obteniendo 0.96 y 0.95 respectivamente.

2.3 Antecedentes Internacionales

A nivel internacional el panomara investigativo es significativamente mucho más extenso en comparación con el nacional, y aunque la mayoría de la publicaciones mantienen el uso de técnicas tradicionales de visión por computador y reconocimiento de patrones, también se percata un especial interés en el uso de técnicas de aprendizaje profundo basadas en CNN para la detección y clasificación de MCs en mamografías digitales.

En [Shin et al., 2015], el autor propone una novedosa técnica de detección de MCs basada en una estructura de clasificación en casacada de tres etapas: **1**. haciendo uso de un clasificador de bosque aleatorio RF para determinar los pixeles donde existe la presencia de MCs de manera individual. El autor utiliza la misma metología propuesta en [Frangi et al., 1998] para extraer características basadas en el uso de la técnica de Vessel para estructuras tubulares, enfocado a estructuras en forma de gota, las cuales representan de manera muy similar a las MCs. Una vez entrenado el clasificador RF, haciendo uso del algoritmo de deslizamiento de ventana (SWM) sobre toda la región del seno que ha sido segmentada previamente como lo sugiere en [Papadopoulos et al., 2002] (umbral adaptativo, dilatación morfológica y etiquetado de componentes conectados (CCL)), se calculan las características construyendo el vector v provenientes de la matriz de valores propios, donde ingresa a al algoritmo RF y clasifican las MCs de manera individual. Dado que la metodología SWM se basa en el corrimiento de la ROI a una distancia d *pixeles* y porcentaje de traslape Θ generalmente no menor al 70 %, múltiples ventanas en un área específica clasifican la presencia de (MCs) de manera redundante, para lo cual, el autor realiza la técnica de supresión no máxima (NMS) con el objetivo de seleccionar una única entidad ROI, que represente la presencia de MCs de manera individual entre las múltiples entidades superpuestas, resultado de la metología SWM. **2**. Con las MCs individuales clasificadas en la aterior etapa, se extraen características de morfología y mediante el uso de la técnica de clasificación máquina de Boltzmann restringida discriminativa (DRBM) se logra mejorar el poder discriminativo. **3**. Finalmente, utilizando una regla estadística basada en la frecuencia de aparición de MCs en un área delimitada, se determina si hay existencia de una agrupación de MCs. La arquitecutra planteada por el autor es evalua en las base de datos MIAS, miniMIAS y (SNUBH-MDB), donde concluye una notable mejoría en

términos de precisión de clasificación para las bases de datos MIAS y SNUBH-MDB, sin embargo, menos evidente para miniMIAS, que tiene una resolución espacial mucho más baja. El autor menciona que el método de clasificación en cascada tiene un rendimiento elevado para imagenes de alta resolución, mientras que para las de baja resolución no tanto, dado a la inestabilidad en el cáculo de la matriz de valores propios junto a las dificultades para capturar diferencias en la morfología de estructuras de pixeles más pequeños.

Por otro lado, son mútiples las publicaciones que están relacionadas con la segmentación de MCs en mamografías digitales. En [Fadil et al., 2019], el autor propone implementar seis de los métodos más usados en la literatura científica para la segmentación de MCs con el objetivo de comparar el rendimiento en el uso de cada una de las técnicas. La eficiencia de los métodos de segmentación es evaluada usando cinco métricas. Los seis métodos de segmentación son: 1. Método de Otsu's, 2. Método de la máxima entropia, 3. Método del umbral iterativo, 4. Método basado en algoritmo genético, 5. Método de segmentación de cuencas hidrográficas y 6. segmentación basada en la transformada de Wavelet. Las cinco métricas son: a. índice de similaridad (SI), b. fracción de superposición adicional (EF), c. sensibilidad, d. especificidad y e. exactitud. El autor reporta que el método de la entropía es más exacto que las otras técnicas. En el caso del algoritmo genético, funciona mejor que los métodos restantes en términos del SI pero con una mayor especificidad. Es importante mencionar que los resultados arrojan un mejor desempeño de la técnica basada en la segmentación de cuencas hidrográficas para imágenes con presencia de tejido fibroglandular, mientras que la técnica de Otsu's es más eficiente para senos grasos. El método de segmentación basado en la transformada de Wavelet y el método del umbral iterativo es adecuado de manera similar para todas las categorías de senos densos. El método basado en el algoritmo genético funciona mejor para senos extremadamente densos, mientras que la técnica de máxima entropía funciona mejor que las anteriores técnicas de manera general para cualquier tipo de densidad mamaria.

Desde otro punto de vista, en la literatura también se reporta investigaciones para la detección automática de MCs sin usar un proceso de aprendizaje de máquina necesariamente. En [Basile et al., 2019], el autor propone una metodología de tres fases para la detección automática de grupos de MCs haciendo uso del repositorio

BCDR y específicamente el uso de las imágenes tipo FFDM, con el objetivo de evitar los procesos más exhaustivos de pre-procesamiento que implica usar imágenes tipo SFM, como lo es: eliminación de ruido perteneciente a la digitalización de la imagen y supresión de artefactos. Las tres fases son: a. pre-procesamiento para resaltar las estructuras mamarías, b. detección de MCs de manera individual y c. detección de grupos de MCs. El ***pre-procesamiento-(a)*** se basa en realizar cinco procesos: a1. Dado que la distribución en la escala de grises de una imagen mamográfica generalmente es estrecha, lo que causa que los detalles sean borrosos, se realiza una mejora del contraste, mediante una modificación del histograma que permita una ampliación en la distribución uniforme de la escala de grises con el objetivo de incrementar el contraste de la imagen y distinguir los detalles. a2. haciendo uso del algoritmo de bola deslizante (the rolling ball algorithm), usado comúnmente en imágenes médicas, se logra remover las variaciones de intensidad del fondo, lo que permite eliminar fondos continuos de manera suaviazada, permitiendo así resaltar las estructuras mamarias. a3. se procede a calcular el gradiente de la imagen mediante el filtro Sobel para detectar los bordes de la imagen. a4. se procede a realizar un suavizado de la imagen mediante un filtro Gaussiano para conservar las estructuras mamarias, y nuevamente en a5. se calcula el filtro Sobel obteniendo como resultado una imagen que contiene los brodes más evidentes de las estructuras de interés de la mama. Una vez obtenida una imagen que resalta la presencia de MCs con respecto al resto de estructuras mamarias, se procede con la ***detección de MCs de manera individual-(b)***, para lo que el autor parte de la consideración de que las MCs son de forma circular y propone dos etapas: b1. utilizar la transformada de Hough con el objetivo de encontrar estructuras morfológicas similares a una forma circular detectando de manera individual las MCs y b2. incrementar el contraste de la imagen pre-procesada mediante un método de dos umbrales para evitar la poca diferencia en imagenes que presentan un tejido glandular denso e imagenes con tejido glandular normal. La ***detección de grupos de MCs-(c)*** se realiza mediante el método de SWM sobre toda la imagen y descartando todas las MCs que se encuentren en el borde del seno, dada su poca importancia radiológica como se menciona en [Sickles, 1986]. En las zonas alejadas del borde del seno, se aplican dos técnicas de agrupamiento para identificar automáticamente el grupo de MCs. Para identificar el k numero de grupos, se aplica el método de agrupamiento aglomerativo jeararquico de Ward's a la distancia euclidiana entre las coordenadas de las MCs de la imagen de manera sucesiva y finalmente para optimizar el proceso se aplica un algoritmo no

jerárquico k-medias. El autor reporta unos resultados experimentales de la metodología propuesta con una sensibilidad en la clasificación del 91.78 %. La revisión realizada por el autor de las investigaciones realizadas de la misma área siguen presentando un nivel inferior de sensibilidad en la clasificación comparado con la metodología propuesta.

2.4 Discusión

La anterior revisión de la literatura a nivel nacional e internacional acerca de las investigaciones más reelevantes realizadas sobre la detección automática de MCs evidencia una diversidad de metodologías propuestas. Se descata que las metodologías estan divididas en dos formas de solución principalmente:

- Metodología basada en el uso de algoritmos de visión por computador para la segmentación y posterior detección de MCs, sin el uso de algoritmos de aprendizaje de máquina.
- Metodología basada en el uso de algoritmos de visión por computador incluyendo el proceso de entrenamiento y clasificación basado en el uso de algoritmos de aprendizaje de máquina.

Por otro lado, se puede observar que sin importar la metodología a usar, generalmente las etapas usadas más importantes son las siguientes:

- **Pre-procesamiento**: El objetivo principal es eliminar cualquier tipo de ruido y artefactos presentes en la imagen mamográfica. Generalmente las imágenes que presentan estos elementos son las mamografías de tipo SFM, mientras que las tipo FFDM no. Algunas veces también es común realizar una mejora de la imagen mediante el ajuste de contraste.
- **Delimitación del seno**: Debido a que la presencia de MCs en un seno tiende a darse en la región retroareolar, se suele delimitar el borde del seno para eliminar el fondo. Por otro lado, dado que en una imagen mamográfica está presente el tejido

pectoral, también es común demarcarlo y eliminarlo. Lo anterior se realiza para evitar inspeccionar regiones donde radiológicamente no representa un interés la aparición de MCs y de otra parte, disminuye el costo computacional al no realizar esta inspección en estas zonas.

- **Segmentación de regiones con presencia de MCs**: Soluciones basadas en visión por comptuador, generalmente deben inspeccionar ROIs para posteriormente extraer características. En la detección automática de MCs, es muy usual que se sugiera utilizar toda la imagen como área de inspección, evitando el uso de algún tipo de algoritmo de segmentación. De la misma manera, muchas investigaciones realizan la segmentación de manera manual (basados en la información previa de su base de datos) o extraer ROIs de la imagen de manera aleatoria sobre una porción de interés de la imagen. Finalmente, se pueden utilizar algoritmos como SWM para inspeccionar de manera automática una región en particular.

- **Extracción de características**: Una vez obtenida las ROIs, se requiere extraer las características para su posterior análisis, cuyo objetivo es poder identificar cuales de éstas representan el mejor poder discriminante entre las clases deseadas a identificar. En sistemas para la detección automática de MCs, es común definir dos clases: **a**. región con presencia de MCs y **b**. región sana (sin presencia de MCs). Algunos autores reportan el interés de no solo indentificar la clase **a**, sino también agrupaciones de MCs, dada la importancia radiológica que éstas representan en relación con un posible CM.

- **Selección de características**: Cuando se ha calculado el paquete de características deseado, es común aplicar diferentes tipos de algoritmos de selección de características cuyo objetivo es poder identificar cuales aportan el mayor poder discriminativo y poder realizar una buena selección de características dependiendo de el uso de éstas.

- **Clasificación**: Una vez se ha seleccionado el conjunto de características que contienen el mayor poder discriminante de las clases, se suele utilizar algoritmos de aprendizaje de máquina para ser entrenados y posteriormente evaluados. Cabe resaltar que en el caso de la detección de grupos de MCs, se realiza generalmente una post-detección sobre la clasificación realizada utilizando métodos basados en visión por computador y algoritmos de agrupamiento. Por

otro lado, esta etapa no necesariamente es llevada a cabo en todos los casos, todo depende de si se desea utilizar procesos basados en el uso de algoritmos de aprendizaje de máquina o no.

Las figuras 2.3, 2.4, 2.5, 2.6 y 2.7 esquematizan los métodos empleados en cada una de las etapas mencionadas anteriormente, a partir de la revisión de la literatura realizada.

Por otra parte, en la tabla 2.1 se relacionan algunos trabajos encontrados en la literatura científica que aportan significativamente en este trabajo de investigación.

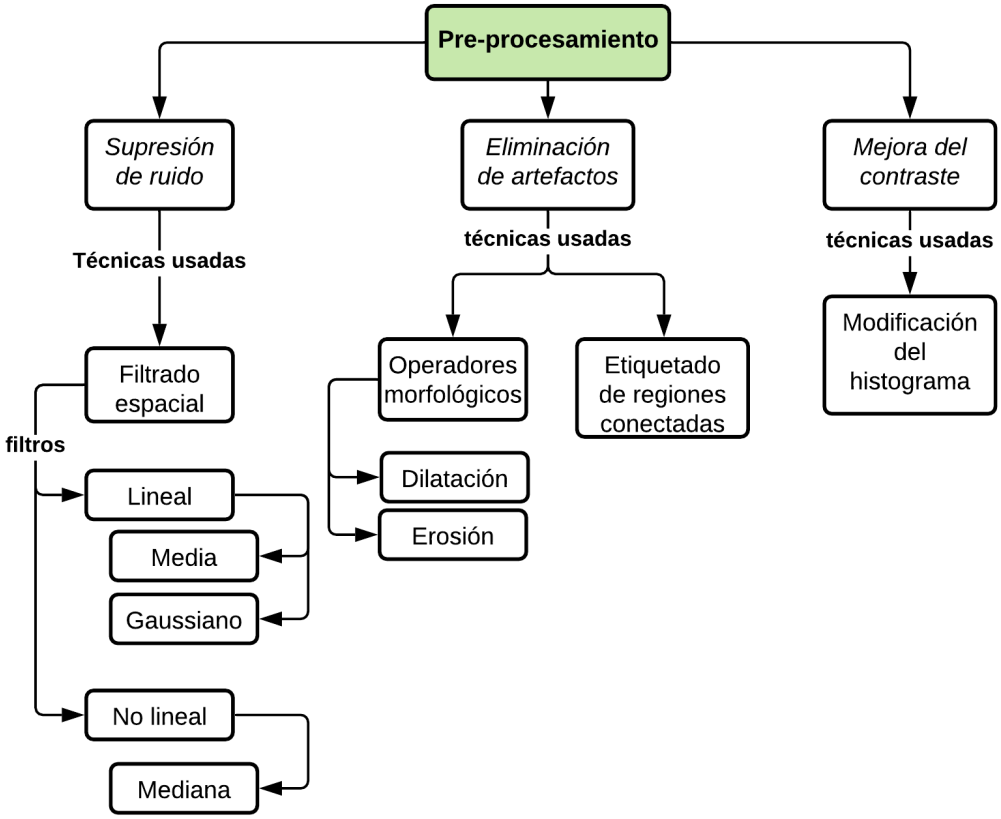


Figura 2.3: Revisión de la etapa de pre-procesamiento en la literatura.

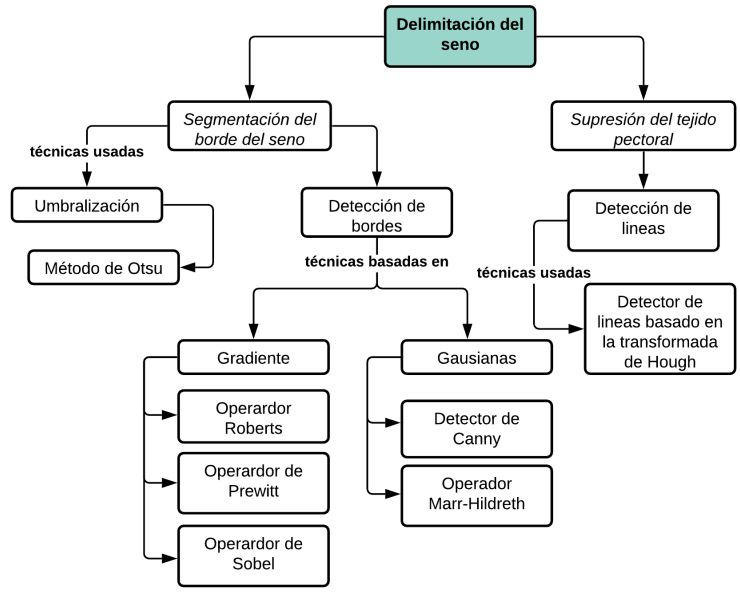


Figura 2.4: Revisión de la etapa de delimitación del seno en la literatura.

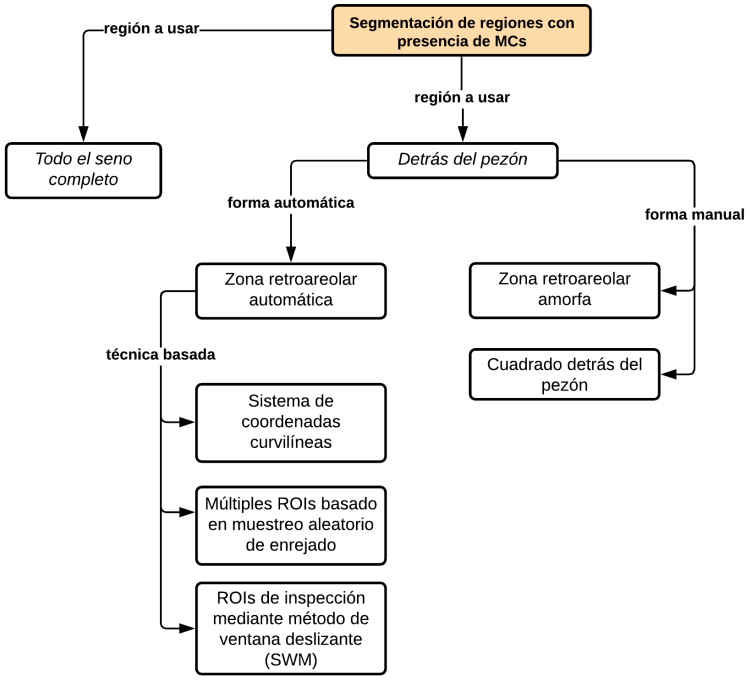


Figura 2.5: Revisión de la etapa de segmentación de regiones con presencia de MCs en la literatura.

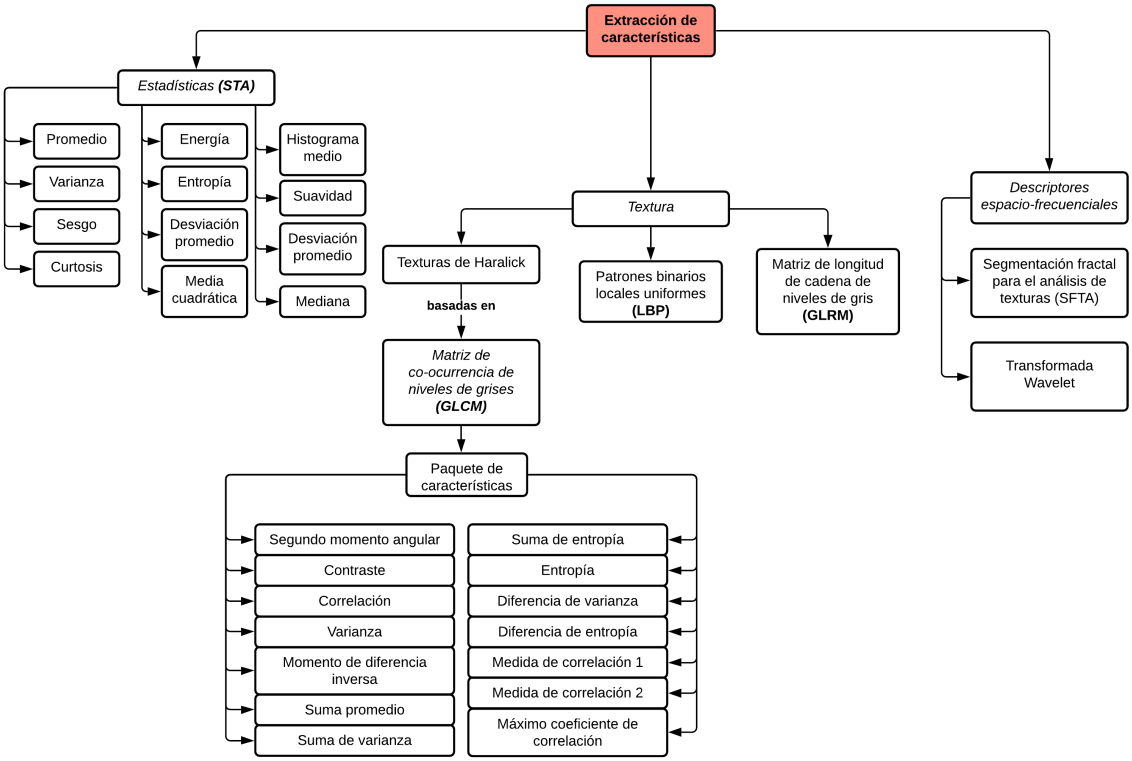


Figura 2.6: Revisión de la etapa de extracción de características en la literatura.

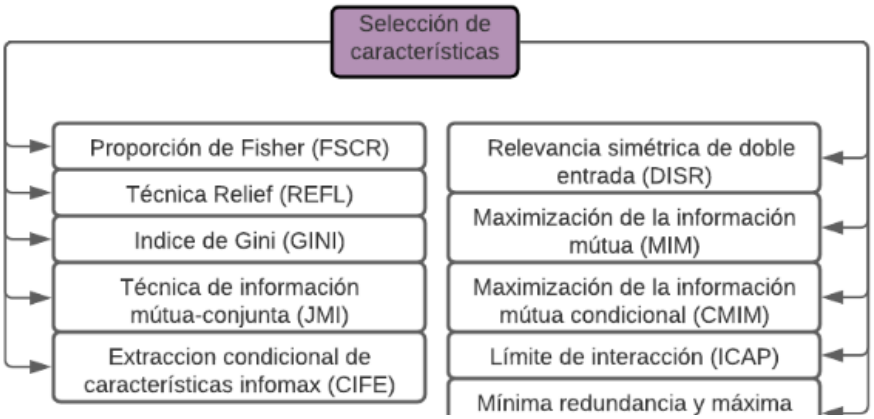


Figura 2.7: Revisión de la etapa de selección de características en la literatura.

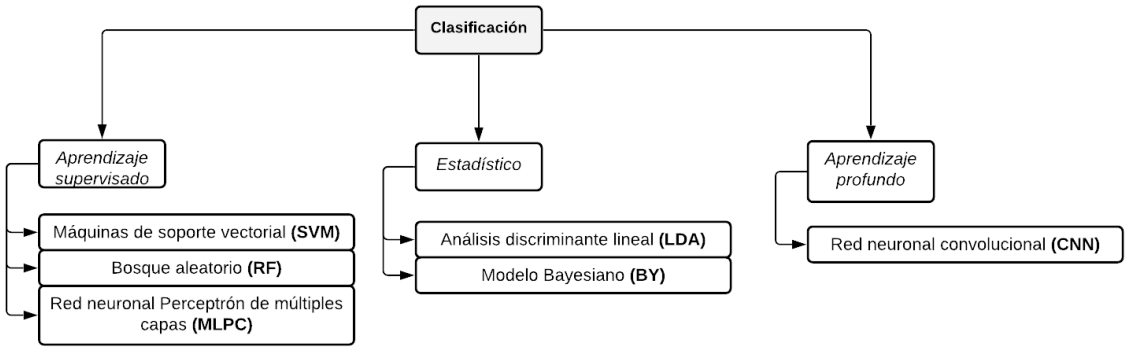


Figura 2.8: Revisión de la etapa de selección de características en la literatura.

2.4.1 Conclusiones de los antecedentes

- Segmentar la mamografía de tal manera que la región de interés a inspeccionar no contenga el fondo ni el tejido pectoral del seno contribuye a reducir la cantidad de falsos positivos en la detección de MCs en una mamografía digital y ser más eficiente computacionalmente.
- La presencia de MCs en la región cercana al borde del seno no representa un interés radiológico en la detección de patologías relacionadas con MCs.
- Dado a que la imagen mamográfica presenta un bajo contraste, es importante realizar una etapa de pre-procesamiento a la imagen mamográfica independientemente si es de tipo SFM o FFDM cuyo objetivo sea mejorar el contraste y permitir diferenciar elementos de interés como las MCs.
- La literatura reporta una etapa de extracción de características basada mayoritariamente en características tipo estadísticas STA y de textura GLCM.
- Las métricas de similitud entre imágenes como los índices Sorensen-Dice y Jaccard como métodos de evaluación del desempeño de la etapa de clasificación, permiten estimar la certeza de las MCs a partir de la percepción visual que se tiene de éstas y el GroundTruth. Estos son considerados los métodos de evaluación más empleados y directos.
- Las métricas más utilizadas para evaluar el resultado de clasificación son el análsis de las curvas ROC, FROC, sensibilidad y especificidad.

Referencia	**Nivel**	**Tipo de imagen mamográfica**	**Base de datos**	**Etapas**						
				Pre-procesamiento	*Delimitación del seno*		*Segmentación*	*Extracción de características*	*Selección de características*	*Clasificación*
					Contorno del seno	Contorno del tejido pectoral				
[Pertuz et al., 2019]	Nacional	▪ SFM ▪ FFDM	BCDR	N.A	▪ Umbralización para FFDM ▪ Método Goodness-of-fit para SFM	Detector de linea basado en la tranformada de Hough	▪ Todo el seno (manual) ▪ ROI cuadrada (manual) ▪ Región retroareolar (automático) ▪ Enrejado muestreado (automático)	▪ STA ▪ GLCM ▪ GLRLM ▪ GRA ▪ SFA	N.A	Regresor logístico
[Vargas et al., 2014]	Nacional	FFDM	EJECALS	Filtro de desenfoque	Operación morfológica (erosión)	N.A	Todo el seno	▪ SFTA ▪ LBP	N.A	SVM
[Rincón et al., 2018]	Nacional	SFM	miniMIAS	N.A	N.A	N.A	Región cuadrada (manual)	▪ DWT ▪ STA ▪ GLCM ▪ LBP	▪ FSCR ▪ REFL ▪ GINI ▪ JMI ▪ CIFE ▪ DISR ▪ MIM ▪ CMIM ▪ ICAP ▪ MRMR	▪ SVM ▪ MLPC ▪ RF ▪ NB ▪ LDA
[Shin et al., 2015]	Internacional	▪ SFM ▪ FFDM	▪ MIAS ▪ miniMIAS ▪ SNUBH-MDB	N.A	▪ Umbral adaptativo ▪ CCL	N.A	▪ SWM ▪ NMS	Técnica de Vessel	N.A	▪ RF ▪ DRBM
[Basile et al., 2019]	Internacional	FFDM	BCDR	▪ Modificación del histograma ▪ Rolling ball algorithm ▪ Filtro Sobel ▪ Filtro Gaussiano	N.A	N.A	Región cuadrada (manual)	Transformada de Hough	N.A	▪ Algoritmo de Ward's ▪ Algoritmo k-medias
[Fadil et al., 2020]	Internacional	FFDM	DDSM	DWT	N.A	N.A	Región cuadrada (manual)	GLCM	N.A	RF
[Ge et al., 2006]	Internacional	▪ FFDM ▪ SFM	Repositorio privado	Filtro de borde caja	▪ Umbralización de Otsu ▪ Conectividad de etiquetado	N.A	Todo el seno (manual)	Características propias de CNN	LDA	▪ LDA ▪ CNN

Tabla 2.1: Antecedentes relevantes.

3. Marco teórico

3.1 Cáncer de mama

El cáncer de seno (mama) se origina cuando las células mamarias comienzan a crecer sin control[1]. Este tipo de cáncer es el que se presenta con más frecuencia en las mujeres. Las células cancerosas del seno normalmente forman un tumor que a menudo se puede observar en una radiografía o se puede palpar como una masa. El cáncer de seno ocurre casi exclusivamente en las mujeres, pero los hombres también lo pueden padecer.

3.1.1 Métodos de diagnostico del cáncer de mama

Los médicos realizan muchas pruebas para detectar o diagnosticar el cáncer de mama. También realizan pruebas para averiguar si el cáncer se ha diseminado a una parte del cuerpo que no sea la mama o los ganglios linfáticos debajo del brazo. Cuando esto sucede, se lo denomina metástasis.

Los tumores no cancerosos de los senos (benignos) son crecimientos anormales, pero no se propagan fuera de los senos. Estos tumores no representan un peligro para la vida, aunque algunos tipos de masas benignas pueden aumentar el riesgo de una mujer de padecer cáncer de seno.

Algunos de los métodos que se utilizan para la detección de cáncer de mama son: biopsia, endoscopia e imagenología.

Una biopsia es la extirpación de una cantidad pequeña de tejido para su examen a través de un microscopio. Otras pruebas pueden indicar la presencia de cáncer, pero solo una biopsia permite formular un diagnóstico definitivo. Luego, un patólogo analiza la(s) muestra(s). Un patólogo es un médico que se especializa en interpretar pruebas de laboratorio y evaluar células, tejidos y órganos para diagnosticar enfermedades. Existen diferentes tipos de biopsias, que se clasifican según la técnica o el tamaño de la aguja

[1]https://www.cancer.org/es/cancer/cancer-de-seno/acerca/que-es-el-cancer-de-seno.html

utilizada para obtener la muestra de tejido.

En los procedimientos endoscópicos, un tubo muy delgado, flexible y liviano adherido a una cámara se inserta a través del pezón y recorre los conductos de la mama en la parte más profunda. Durante el procedimiento se pueden extraer muestras de tejido y líquidos.

Las pruebas por imágenes se enfocan en partes del interior del cuerpo. Se pueden realizar las siguientes pruebas por imágenes en la mama para saber más acerca de un área sospechosa encontrada en la mama durante un examen de detección.

3.1.1.1 Imagenología

Se presentan tres tipos de pruebas con imágenes:

- Mamografía de diagnóstico. La mamografía de diagnóstico es similar a la mamografía de detección, salvo que se toman más imágenes de la mama. Por lo general, se utiliza cuando la mujer experimenta signos, como una masa nueva o secreción del pezón. La mamografía de diagnóstico también puede utilizarse si en una mamografía de detección se encuentra algo sospechoso.
- Ecografía. La ecografía utiliza ondas de sonido para crear una imagen del tejido mamario. Una ecografía puede distinguir entre una masa sólida, que puede ser cáncer, y un quiste lleno de líquido, que generalmente no es canceroso.
- Resonancia magnética (RM). Una RM usa campos magnéticos, en lugar de rayos X, para producir imágenes detalladas del cuerpo. Se administra un tinte especial, llamado medio de contraste, antes de la exploración para ayudar a crear una imagen clara del posible cáncer. Este tinte se inyecta en una vena del paciente. Una RM de la mama también es una opción de detección, junto con una mamografía, en algunas mujeres con un riesgo muy elevado de desarrollar cáncer de mama. Por último, la RM se puede utilizar como método de vigilancia después de un diagnóstico y tratamiento de cáncer de mama.[2]

[2]https://www.cancer.net/

3.2 Mamografías digitales

La Mamografía digital, también llamada mamografía digital de campo completo (MDCC), es un sistema de mamografía en el que la película de rayos X es reemplazada por sistemas electrónicos que transforman los rayos X en imágenes mamográficas de las mamas. Estos sistemas son similares a los que tienen las cámaras digitales y su eficiencia permite obtener mejores fotografías con una dosis más baja de radiación. Estas imágenes de las mamas se transfieren a una computadora para su revisión por un radiólogo y para su almacenamiento a largo plazo. La experiencia del paciente durante una mamografía digital es similar a la de una mamografía convencional.[3]

La mamografía puede detectar a menudo el cáncer tempranamente, cuando aún este es muy pequeño o inclusive no hay presencia de algún abultamiento. En esta etapa el cáncer es más fácil de tratar.

3.2.1 Calcificaciones

Las calcificaciones mamarias más comunes son distróficas, se producen y algunas veces incluso están asociados, durante diferentes procesos patológicos: inflamación, infección, tumor benigno o maligno [Henrot et al., 2014].

En muchas mujeres que se someten a una mamografía se encuentran calcificaciones en el seno. La gran mayoría de estas calcificaciones son claramente benignas, por lo que causan poco problema en el diagnóstico. También se observan calcificaciones típicas en los cánceres de mama, mostrándose en grupos de partículas diminutas caracterizadas por formas lineales, curvilíneas o ramificadas finas [Sickles, 1986]. Estas calcificaciones malignas, aunque no suelen ser palpables, suelen ser tan fáciles de identificar y evaluar como las benignas.

[3]https://www.radiologyinfo.org/

3.2.1.1 Tipos de calcificaciones

Existen las macro calcificaciones y las microcalcificaciones, este ultimo se refiere a las calcificaciones cuyo diámetro es inferior a 1 mm, sabiendo que los mamógrafos de resolución espacial actuales hacen que los objetos pequeños se detecten sin aumento para un tamaño que oscila entre 100 y 200 μm [Henrot et al., 2014].

3.2.1.2 Benignidad y malignidad

A grandes rasgos, las microcalcificaciones grandes de aproximadamente un milímetro son más a menudo benignas que aquellas cuyo tamaño es inferior a 0,5 mm. Sin embargo, hay excepciones, ya que las calcificaciones gruesas heterogéneas o distróficas de más de un milímetro pueden asociarse a lesiones malignas.

Un grupo de 10 o más microcalcificaciones en un volumen inferior a 1 cm3 es más sospechoso que un grupo de 5 de idéntica morfología.

El lugar de formación o acumulación de las calcificaciones puede determinar su morfología:

- redondas o punteadas, de tamaño normal o dilatadas, en los acinos lobulares, más a menudo asociadas a lesiones benignas o a carcinomas in situ de bajo grado.
- lineal en la luz de un conducto pequeño por calcificación de una zona de necrosis asociada a un carcinoma ductal in situ de alto grado o por calcificaciones en forma de varilla de las secreciones acumuladas en un conducto mayor como consecuencia de una galactoforitis.
- gruesa en el tejido conectivo de un adenofibroma o en la reacción del estroma de un carcinoma invasivo [Henrot et al., 2014].

3.3 Visión artificial

El campo de la visión artificial se enfoca en describir el mundo que se observa en una o muchas imágenes y posteriormente, utilizando distintas técnicas matemáticas, reconstruir sus propiedades tales como forma, iluminación y distribución de color [Szeliski, 2010].

3.3.1 Técnicas de preprocesamiento de imágenes

Generalmente, una de las primeras etapas de un sistema de visión artificial es el preprocesamiento, en esta se descarta la información no relevante de la imagen y, además, se realzan las propiedades que serán de mayor utilidad para las etapas posteriores.

3.3.1.1 Mejora del contraste

El contraste es una medida de preservación de detalles en una imagen. Un contraste alto permite diferenciar objetos que contienen distintos niveles de grises, mientras que uno bajo dificulta dicha tarea.

El contraste se puede asociar a la distribución de grises en el histograma global de una imagen tal y como se observa en la figura 3.1.

3.3.1.1.1. Transformada exponencial

La transformada exponencial busca modificar el rango de niveles de gris de la imagen de entrada siguiendo la ecuación 3.1, alterando a su vez el contraste de la misma.

$$g(x, y) = c \cdot I(x, y)^{\gamma} \tag{3.1}$$

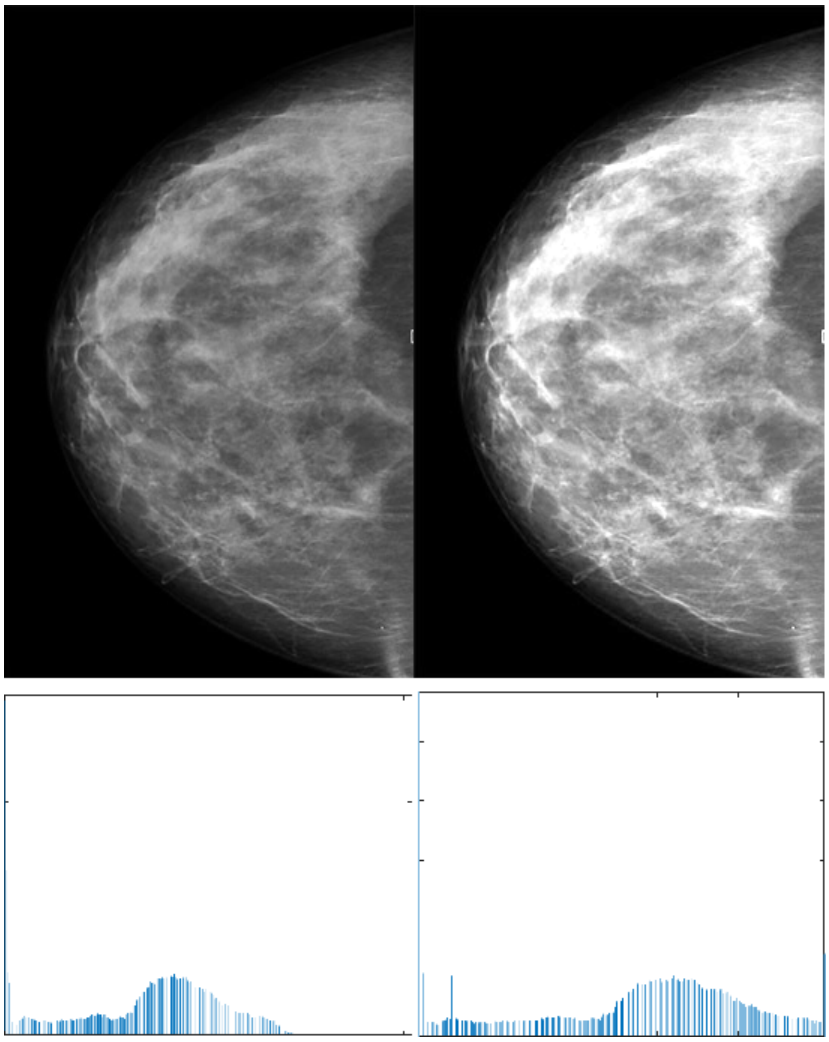


Figura 3.1: Uso incompleto del rango dinámico / Uso de todo el rango dinámico de la imagen [Fuente: Elaboración propia].

Para distintos valores de γ, la imagen sufrirá alteraciones de la siguiente forma:

- Si $\gamma > 1$ Se expanden las franjas de gris claro en la imagen de salida.
- Si $\gamma = 1$ No hay transformación.
- Si $\gamma < 1$ Se expanden las franjas de gris oscuro en la imagen de salida.

Cabe resaltar que para valores muy bajos de γ la imagen sufrirá una pérdida de contraste.

3.3.1.2 Técnicas para supresión de artefactos

Un artefacto o artificio se define como una distorsión, adición o error en una imagen que no tiene correlato en el sujeto o región anatómica estudiada. Como término, deriva de las palabras latinas artis (artificial) y actum (efecto), y refiere a un efecto artificial que altera la calidad y fidelidad de una imagen, pudiendo encubrir una patología o crear hallazgos falsos [Sartori et al., 2015].

Existen múltiples técnicas que se especializan en la detección de regiones, las cuales pueden ser utilizadas para encontrar artefactos dentro de las imágenes.

3.3.1.2.1. Etiquetado de componentes conectados

El CCL es un método utilizado ampliamente en la visión por computador, se basa en el uso de vecindades para encontrar regiones pertenecientes a objetos dentro de una imagen.

Existen distintas vecindades, entre las más usadas se encuentran:

- 4-Vecindad: toma en cuenta como vecinos a los píxeles superiores y laterales.
- 8-Vecindad: toma en cuenta como vecinos a los píxeles superiores, laterales y diagonales.

Una vez seleccionada la vecindad a utilizar, se procede a implementar el algoritmo seleccionado para el etiquetado. Existen una gran variedad de estos, sin embargo se basan en un principio básico de dos pasos:

1. Etiquetado de regiones conexas.
2. Localización de las ventanas de los objetos.

Para el primer paso se puede realizar un recorrido de izquierda a derecha y de arriba hacia abajo, encontrando distintas regiones y presentando equivalencias cuando un píxel puede pertenecer a más de una región.

Luego, se realiza un segundo recorrido, resolviendo las equivalencias que se presenten en el primer recorrido y definiendo así, las regiones encontradas. Este procedimiento se ilustra con un ejemplo en la figura 3.2.

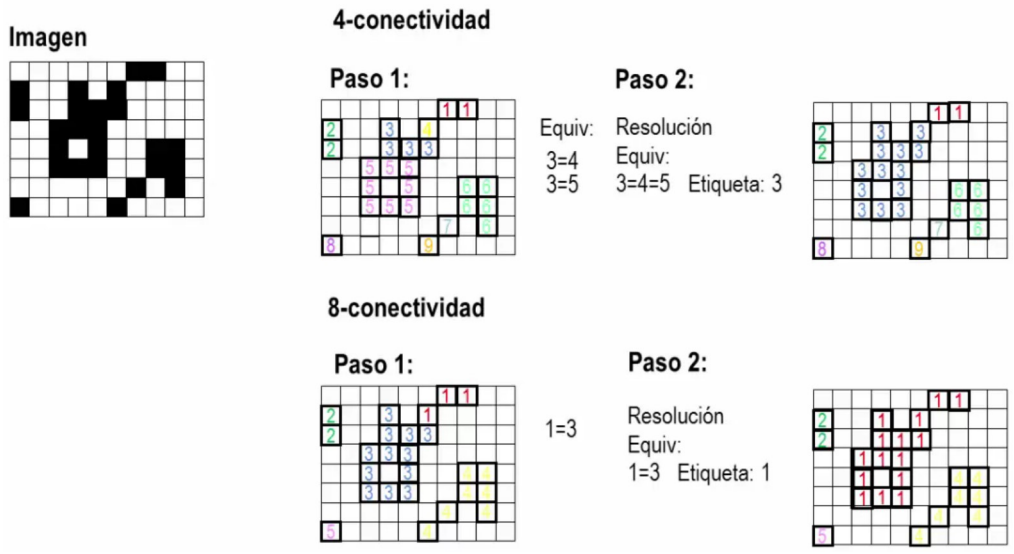


Figura 3.2: Ilustración del primer paso del algoritmo de CCL [Fuente: Componentes conexas, Curso: Detección de objetos, Universidad Autónoma de Barcelona].

Por último, para el segundo paso, simplemente se localizan los píxeles de los extremos inferior y superior de cada región, ubicando así la ventana donde se encuentra el objeto.

3.3.2 Métricas para evaluar la calidad de una imagen

Durante el preprocesamiento que se aplica a las imágenes, se pueden apreciar cambios a simple vista, con los cuales se puede determinar si se presenta una mejora con respecto a la anterior. Sin embargo, en ocasiones los cambios son tan pequeños que se vuelven imperceptibles al ojo humano, por esta razón, existen ciertas métricas que miden la calidad de las imágenes con respecto a diferentes propiedades de las mismas.

3.3.2.1 Frecuencia espacial

Esta métrica es un índice de calidad basado en el cálculo de gradientes verticales (CF) y horizontales de la imagen (RF). Los gradientes también son denominados frecuencia espacial vertical y horizontal. Esta métrica permite medir el detalle y textura de una imagen. Se calcula a partir de:

$$SF = \sqrt{RF^2 + CF^2} \tag{3.2}$$

La frecuencia de las filas (RF) y columnas (CF), son definidos como:

$$RF = \sqrt{\sum_{i=1}^{M}\sum_{j=2}^{N}(I(i,j) - I(i,j-1))^2} \tag{3.3}$$

$$CF = \sqrt{\sum_{i=2}^{M}\sum_{j=1}^{N}(I(i,j) - I(i-1,j))^2} \tag{3.4}$$

A mayor valor de SF en una imagen, implica que ésta es sensible a la percepción humana de acuerdo con el sistema visual humano, evidenciando la riqueza en bordes y texturas.

3.3.2.2 Gradiente promedio

Esta métrica, presenta la información de detalle y textura en una imagen, permitiendo definir la nitidez. Un gran valor de gradiente implica que la imagen

presenta mayor información de textura. Se define como:

$$AG = \frac{1}{M \cdot N} \sum_{i=1}^{M-1} \sum_{j=1}^{N-1} \sqrt{\frac{1}{2} \cdot ((I(i,j) - I(i+1,j))^2 + (I(i,j) - I(i,j+1))^2} \quad (3.5)$$

3.3.2.3 Desviación estándar

Esta métrica basada en un concepto estadístico refleja la distribución y contraste de la imagen. Se define como:

$$STD = \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} (I(i,j) - \mu)^2} \quad (3.6)$$

Donde:

- M, N son las dimensiones de la imagen, filas y columnas.
- I es la imagen en su representación matricial.
- i,j son las posiciones de la fila y la columna de un píxel en la posición iésima, jotaésima.
- μ es el valor medio.

Las regiones que presentan alto contraste son atractivas para el sistema visual humano. Un valor grande de STD refleja un alto de contraste.

3.3.3 Técnicas de segmentación de imágenes

En muchas ocasiones, las imágenes contienen información no deseada o que no es de utilidad para detectar el objeto de interés. Por esta razón, se utilizan distintos métodos

para la segmentación de la ROI.

3.3.3.1 Técnicas basadas en la binarización

Las imágenes binarias o binarizadas se pueden interpretar como imágenes en escala de grises que cuentan con solo 2 niveles, por lo general, se tiene que para el negro es el 0 y para el blanco es el 1.

3.3.3.1.1. Método de Otsu

El método de Otsu es uno de los más exitosos para la umbralización de imágenes [Liu and Yu, 2009]. Consiste en encontrar uno o más umbrales óptimos para así realizar la segmentación de los objetos y el fondo de la imagen.

Para encontrar estos umbrales, Otsu se basa en el histograma global de la imagen, calcula la dispersión de los niveles de gris y describe la varianza entre clases con la ecuación 3.7

$$\sigma_B^2 = \omega_1 \cdot (\mu_1 - \mu_T)^2 + \omega_2 \cdot (\mu_2 - \mu_T)^2 \tag{3.7}$$

Donde:

- ω_1 es la sumatoria de las intensidades de los píxeles de la clase 1.
- ω_2 es la sumatoria de las intensidades de los píxeles de la clase 2.
- μ_1 es la intensidad media de los píxeles de la clase 1.
- μ_2 es la intensidad media de los píxeles de la clase 2.
- μ_T es la intensidad media de toda la imagen.

El umbral óptimo se encuentra maximizando $\sigma_B^2(t)$ de la ecuación 3.7.

Una vez encontrados los umbrales óptimos, se realiza una cuantificación de la imagen dependiendo de cuantos niveles se requieran, para el caso binario, se realiza el etiquetado de acuerdo a la ecuación 3.8.

$$g(x,y) = \begin{cases} 1 \Leftrightarrow f(x,y) > T \\ 0 \Leftrightarrow f(x,y) \leq T \end{cases} \tag{3.8}$$

3.3.3.1.2. Umbral por histograma del gradiente

Esta técnica se basa principalmente en calcular el umbral a partir de la aproximación exponencial del histograma de la magnitud del gradiente. Permite calcular un valor de umbral τ el cual puede ser ajustado de manera proporcional para brindar una aproximación en la umbralización.

La figura 3.3 es la representación gráfica de la técnica que se obtiene a partir de las ecuaciones 3.9 y 3.10

$$n = n_0 \cdot e^{\frac{-g}{\tau}} \tag{3.9}$$

$$\tau = \frac{g_2 - g_1}{ln(\frac{n_1}{n_2})} \tag{3.10}$$

Como se puede observar en la figura 3.3, las coordenadas *Pico 1* y *Pico 2* son los valores de (g_1, n_1) y (g_2, n_2) respectivamente. El valor de τ corresponde al umbral encontrado que es justo el corte de la recta tangente de la exponencial de 3.9 en el eje de los valores de gradiente g.

Para desarrollar el algoritmo se siguen los siguientes pasos:

- Calcular el valor absoluto del gradiente total de la imagen de entrada: Gradiente

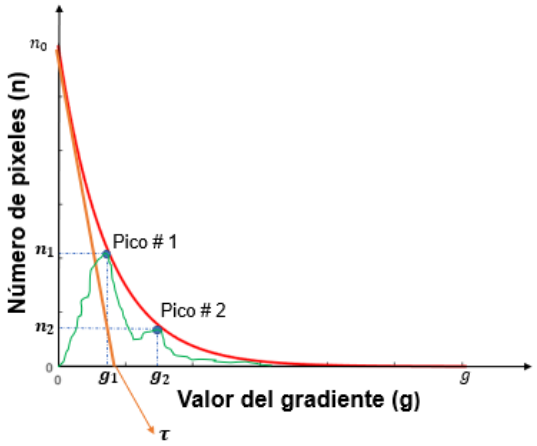


Figura 3.3: Representación gráfica de la técnica Umbral por el Histograma del Gradiente.

Vertical + Gradiente Horizontal.

- Obtener el histograma del gradiente total de la imagen.
- Identificar el valor n_1 y n_2 sobre el gradiente y así sus correspondientes g_1 y g_2.
- Calcular el valor de τ usando 3.10.
- Usando las ecuaciones 3.9 y 3.10, calcular n_0 como se presenta en 3.11

$$n_0 = \frac{n_1 + n_2}{e^{(\frac{-g_1}{\tau})} + e^{(\frac{-g_2}{\tau})}} \tag{3.11}$$

- Trazar la recta tangente con pendiente negativa entre los puntos $(0, n_0)$ y $(\tau, 0)$ para corroborar visualmente el valor del umbral τ.

3.3.3.2 Segmentación compleja de planos de bits

La Segmentación compleja de planos de bits (BPCS) consiste en descomponer una imagen en una serie imágenes binarias, llamadas planos de bits. Estos corresponden a cada posición de los valores de los píxeles, así, el primer plano de bit es una imagen binaria que consiste en los bits menos significativos de cada píxel de la imagen (LSB). El siguiente plano superior consiste en el segundo bit de cada píxel y así sucesivamente [Padrón-Godínez et al., 2008]. En el caso de una imagen de 8 bits, el grupo de imágenes obtenidas es desde el plano 0 hasta el plano 7

3.3.3.3 Métricas para evaluar la segmentación de una imagen

Con el fin de conocer el nivel de calidad de la segmentación realizada, es común establecer una máscara de ground truth, la cual corresponde a una segmentación ideal del objeto en cuestión. Así, una vez obtenida esta máscara, se pueden aplicar distintas métricas que ayudan a conocer que tan bien segmentada se encuentra la ROI.

3.3.3.3.1. Exactitud

Esta métrica mide la proporción de píxeles correctamente identificados respecto al total de píxeles en la imagen. Está dada por la ecuación 3.12.

$$Exactitud = \frac{VP + VN}{VP + FP + FN + VN} \tag{3.12}$$

Donde:

- VP son los verdaderos positivos
- VN son los verdaderos negativos
- FN son los falsos negativos
- FP son los falsos positivos

3.3.3.3.2. Precisión

Esta métrica se encarga de medir la proporción de píxeles correctamente identificados como positivos del total de datos identificados como positivos. Está dada por la ecuación 3.13. Usualmente, se prefiere esta métrica cuando se desea una mayor

confiabilidad en los verdaderos positivos.

$$Precision = \frac{VP}{VP + FP} \tag{3.13}$$

3.3.3.3.3. Sensibilidad

Esta métrica se usa cuando los falsos negativos son inaceptables. También, puede considerarse como la exactitud de los casos positivos y está dada por la ecuación 3.14.

$$Sensibilidad = \frac{VP}{VP + FN} \tag{3.14}$$

3.3.3.3.4. Índice F1

Está métrica corresponde a la media armónica de los píxeles, se combinan la sensibilidad y la precisión en una sola métrica. Está dado por la ecuación 3.15

$$F1Score = \frac{2 \cdot Precision \cdot Sensibilidad}{Precision + Sensibilidad} \tag{3.15}$$

Se utiliza cuando tienen más importancia los falsos negativos y los falsos positivos.

3.3.3.3.5. Coeficiente de correlación de Matthews

Esta métrica no depende de la elección de la clase positiva, además, tiene en cuenta los falsos positivos y los falsos negativos. Es la correlación entre la segmentación real y

la obtenida. La ecuación esta dada en 3.16

$$MCC = \frac{(VP \cdot VN) - (FP \cdot FN)}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}} \tag{3.16}$$

3.3.3.3.6. Especificidad

Esta métrica se usa cuando los falsos positivos son inaceptables, puede considerarse como la exactitud en los casos negativos y está dada por la ecuación 3.17

$$Especificidad = \frac{VN}{VN + FP} \tag{3.17}$$

3.3.3.3.7. Coeficiente de Sorensen-Dice

El coeficiente de Sorensen-Dice es utilizada para encontrar la similitud entre dos muestras. Para imágenes, se requieren de dos imágenes binarias: imagen de referencia e imagen a comparar.

En 3.18, A y B representan la imagen de referencia e imagen a comparar respectivamente. Por otro lado $\bigcap$ es la operación intersección. Esta métrica arroja un resultado entre 0 y 1, donde 1 significa que la similitud entre la imagen a comparar y la imagen de referencia es muy alta, mientras que el valor 0 representa que la similitud entre ambas imágenes es nula.

$$Sorensen - Dice = \frac{2\,|A \bigcap B|}{|A| + |B|} \tag{3.18}$$

También se puede expresar como la ecuación 3.19

$$Sorensen - Dice = \frac{2 \cdot VP}{(VP + FP)(VP + FN)} \tag{3.19}$$

3.3.3.3.8. Índice de Jaccard

El índice de Jaccard también es utilizada para encontrar el nivel de similitud entre dos muestras. Como en el coeficiente de Sorensen-Dice, para imágenes se requieren de dos imágenes binarias: imagen de referencia e imagen a comparar.

En 3.20, A y B representan la imagen de referencia e imagen a comparar respectivamente. Por otro lado $\bigcup$ y $\bigcap$ son las operaciones unión e intersección. Esta métrica arroja un resultado entre 0 y 1, donde 1 significa que la similitud entre la imagen a comparar y la imagen de referencia es muy alta, mientras que el valor 0 representa que la similitud entre ambas imágenes es nula.

$$Jaccard = \frac{|A \bigcap B|}{|A \bigcup B|} \tag{3.20}$$

3.3.4 Extracción de características en las imágenes

La gran mayoría de sistemas de visión artificial consisten en interpretar una imagen del mundo real por medio de sus propiedades. Por lo general, los objetos que pertenecen a una misma clase, comparten ciertas características como forma, tamaño, color, textura, etc. Gracias a esto, se han creado una variedad de técnicas para representar y describir estas características.

Usualmente, se buscan características que sean similares para objetos que pertenezcan a la misma categoría y diferentes entre categorías. También, se prefieren características que sean invariantes a operaciones tales como la traslación, rotación, escala y otro tipo de distorsiones, esto debido a que las imágenes nunca suelen presentarse con la misma perspectiva del objeto.

3.3.4.1 Características estadísticas de primer orden

Para hallar las características estadísticas de una imagen, se extraen distintos momentos a partir del histograma global de escala de grises.

Los cuatro primeros momentos estadísticos son:

3.3.4.1.1. Promedio

El promedio es la suma de todos los datos dividido por la cantidad de muestras que se tienen. Se puede expresar por la siguiente ecuación:

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} \tag{3.21}$$

3.3.4.1.2. Desviación estándar

La desviación estándar es un promedio de las desviaciones individuales de cada dato con respecto a la media de una distribución. En otras palabras, la desviación estándar mide el grado de dispersión. Primero, se calcula la diferencia entre cada valor del conjunto de datos y la media del conjunto de datos. Luego, se suman para encontrar el total. Por último, se divide el total por la cantidad de datos para llegar a un promedio de las distancias entre cada dato individual y la media.

La desviación estándar normalmente se representa por la ecuación 3.22:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N}} \tag{3.22}$$

3.3.4.1.3. Asimetría

Miden la mayor o menor simetría de la distribución. Existen dos medidas de este tipo:

Índice de simetría de Pearson:

$$P_1 = \frac{\bar{X} - Mo}{\sigma} \tag{3.23}$$

Índice de simetría de Fisher:

$$F_1 = \frac{1}{N} \cdot \frac{\sum_{i=1}^{n} (X_i - \bar{X})^3 \cdot P_i}{\sigma^3} \tag{3.24}$$

Si el histograma es simétrico, ambos índices dan 0, mientras que si es asimétrica a la derecha, ambos son positivos; y si es asimétrica a la izquierda, ambos índices son negativos.

3.3.4.1.4. Curtosis

Miden la mayor o menor concentración de datos alrededor de la media. Se suele medir con el coeficiente de curtosis:

$$C_1 = \frac{1}{N} \cdot \frac{\sum_{i=1}^{n} (X_i - \bar{X})^4 \cdot P_i}{\sigma^4} - 3 \tag{3.25}$$

- Si este coeficiente es nulo o 0, la distribución se dice normal (similar a la distribución normal de Gauss) y recibe el nombre de mesocúrtica.
- Si el coeficiente es positivo, la distribución se llama leptocúrtica, más puntiaguda que la anterior. Hay una mayor concentración de los datos al rededor de la media.
- Si el coeficiente es negativo, la distribución se llama platicúrtica y existe menor concentración de datos al rededor de la media.

3.3.4.1.5. Entropía

La entropía es una medida estadística de aleatoriedad que se puede utilizar para

caracterizar la textura de la imagen de entrada. La entropía se define como

$$-\Sigma(p \cdot log_2(p)) \tag{3.26}$$

Donde p contiene los recuentos de histograma normalizados.

3.3.4.1.6. Energía

La energía que porta una señal o imagen es una medida muy útil. Se define como la suma de los cuadrados de los valores de la señal:

$$E = \int_{t_1}^{t_2} |x(t)|^2 dt \tag{3.27}$$

3.3.4.1.7. Mediana
La mediana estadística representa el valor de la variable de posición central en un conjunto de datos ordenados.

3.3.4.2 Características basadas en textura

La textura en una imagen se entiende como la estructura de la capa superficial de un material, una foto con una la textura muy resaltada, confiere realismo a la imagen. La textura, junto con el tono y la forma, transforman los motivos planos en imágenes con fuerte sensación tridimensional.

La mayoría de las veces, cuando se tiene un área pequeña de una imagen donde existen pequeños cambios de niveles de grises, se dice que la característica predominante es el tono. Por otro lado, cuando en esta área, se presentan cambios fuertes en los niveles de gris, la característica que predominará es la textura [Haralick et al., 1973].

3.3.4.2.1. Características de textura de Haralick

Uno de las métodos más utilizados para extraer características de textura es el método propuesto por Haralick en [Haralick et al., 1973]. Para desarrollar este método, se necesita encontrar primero la GLCM.

La GLCM es una matriz que registra las veces que se repiten determinados pares adyacentes de píxeles en una imagen (correspondientes a las coordenadas de los índices en la GLCM). El tamaño de la matriz resultante dependerá del máximo nivel de gris, para una imagen de 8 bits, se tendría una matriz de 255 x 255.

Usualmente, se calculan cuatro GLCMs a una distancia determinada en píxeles y con un ángulo de la siguiente forma:

- De izquierda a derecha ⇒ 0°.
- De la esquina superior derecha a la inferior izquierda ⇒ 45°.
- De arriba a abajo ⇒ 90°.
- De la esquina superior izquierda a la inferior derecha ⇒ 135°.

El método se ilustra en la figura 3.4 para el cálculo de una GLCM a 0° con una distancia de 1 píxel.

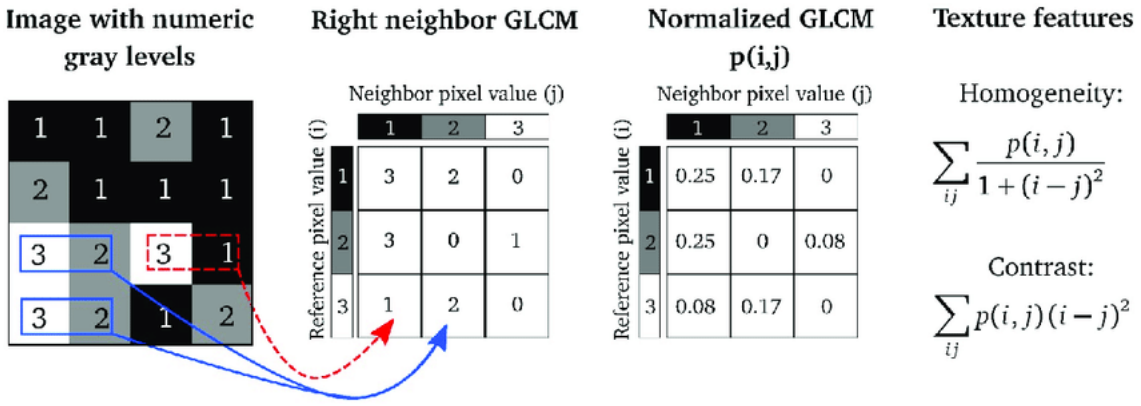


Figura 3.4: Funcionamiento del descriptor de Haralick. Fuente: [Löfstedt et al., 2019].

Se pueden extraer hasta 14 características con el método de Haralick y son las siguientes:

1. Segundo momento angular.

$$f_1 = \sum_i \sum_j p(i,j)^2 \tag{3.28}$$

2. Contraste.

$$f_2 = \sum_{n=0}^{N_{g-1}} n^2 \cdot \left\{ \begin{pmatrix} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \end{pmatrix} \right\} \tag{3.29}$$
$$|i-j| = n$$

3. Correlación.

$$f_3 = \frac{\sum_i \sum_j (ij) \cdot p(i,j) - \mu_x \cdot \mu_y}{\sigma_x \cdot \sigma_y} \tag{3.30}$$

Donde μ_x, μ_y, σ_x y σ_y son los promedios y las desviaciones estándar de p_x y p_y.

4. Varianza.

$$f_4 = \sum_i \sum_j (i-\mu)^2 \cdot p(i,j) \tag{3.31}$$

5. Homogeneidad.

$$f_5 = \sum_i \sum_j \frac{1}{1+(i-j)^2} \cdot p(i,j) \tag{3.32}$$

6. Suma promedio.

$$f_6 = \sum_{i=2}^{2N_g} i p_{x+y}(i) \tag{3.33}$$

7. Suma de la varianza.

$$f_7 = \sum_{i=2}^{2N_g} (i-f_8)^2 \cdot p_{x+y}(i) \tag{3.34}$$

8. Suma de la entropía.

$$f_8 = -\sum_{i=2}^{2N_g} p_{x+y}(i) \cdot log(p_{x+y}(i)) \tag{3.35}$$

9. Entropía.

$$f_9 = -\sum_i \sum_j p(i,j) \cdot log(p(i,j)) \tag{3.36}$$

10. Varianza diferencial.

$$f_{10} = f_4(p_{x-y}) \tag{3.37}$$

11. Entropía diferencial.

$$f_{11} = -\sum_{i=0}^{N_{g-1}} p_{x-y}(i) \cdot log(p_{x-y}(i)) \tag{3.38}$$

12. Medida de correlación I.

$$f_{12} = \frac{HXY - HXY1}{max(HX, HY)} \tag{3.39}$$

13. Medida de correlación II.

$$f_{13} = \sqrt{(1 - exp[-2 \cdot (HXY2 - HXY)])} \tag{3.40}$$

Donde:

- $HXY = -\sum_i \sum_j p(i,j) \cdot log(p(i,j))$
- HX y HY son las entropías de p_x y p_y
- $HXY1 = -\sum_i \sum_j p(i,j) \cdot log(p_x(i) \cdot p_y(j))$
- $HXY2 = -\sum_i \sum_j p_x(i) \cdot p_y(j) \cdot log(p_x(i) \cdot p_y(j))$

14. Correlación máxima de coeficientes.

$$f_{14} = \sqrt{(segundo\ valor\ propio\ de\ Q)} \tag{3.41}$$

Donde $Q(i,j)) = \sum_k \frac{p(i,k) \cdot p(j,k)}{p_x(i) \cdot p_y(k)}$

Donde:

- $p(i,j)$ posición de un dato en una matriz espacial de tonos de gris.
- $p_x(i)$ iésimo dato en la matriz de probabilidad marginal obtenida de la sumatoria de las filas de p(i,j) = $\sum_{j=1} N_g P(i,j)$.
- N_g Número de niveles de grises distintos en la imagen cuantificada.
- $\sum_i$ y $\sum_j$ son $\sum_{i=1}^{N_g}$ y $\sum_{j=1}^{N_g}$ respectivamente.
- $p_y(j) = \sum_{i=1}^{N_g} p(i,j)$
- $p_{x+y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)$ con $i + j = k$ y $k = 2, 3, \ldots, 2N_g$
- $p_{x-y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)$ con $|i - j| = k$ y $k = 0, 1, \ldots, N_g - 1$

3.3.4.2.2. Características de textura utilizando Local Binary Pattern

El LBP es una técnica de gran éxito en la visión artificial y el reconocimiento de patrones [Guo et al., 2010].

Dado un píxel en una imagen, se calculará un nuevo valor de píxel denominado código LBP realizando una comparación con sus vecinos, siguiendo la ecuación 3.42.

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c) \cdot 2^p, s(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \tag{3.42}$$

Donde:

- g_c es el valor de píxel central.

- g_p es el valor de cada píxel vecino.
- P es el numero total de vecinos.
- R es el radio de el vecindario que se escoja.

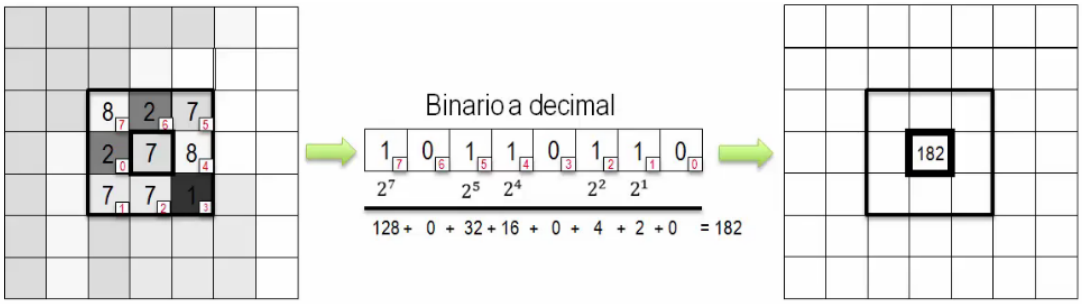


Figura 3.5: Ilustración del funcionamiento básico del algoritmo LPB. [Fuente: Local Binary Patterns, Curso: Detección de objetos, Universidad Autónoma de Barcelona]

Una vez se tenga la nueva imagen con los códigos LBP calculados, se procede a hallar su histograma global. Con esta herramienta se tendrán N características, donde N dependerá de la cantidad de bits de la imagen.

Para una imagen de 8 bits se tendrán 255 niveles que pertenecerán al histograma.

3.3.4.2.3. Local Binary Pattern Uniform

La diferencia con el método anterior es simplemente que de los N niveles de grises se seleccionan los que tengan un patrón uniforme, es decir, que presenten solo dos variaciones entre 1 y 0 en su código binario. Para la imagen de 8 bits, se tienen 59 posibilidades de patrón uniforme, reduciendo así la longitud del histograma a $\frac{1}{4}$ de su tamaño original, disminuyendo también el tiempo de cómputo requerido por cada imagen.

Por otro lado, cabe destacar que esta transformación es invariante frente a cambios de monotónicos de niveles de gris y a traslaciones en la imagen.

3.3.4.2.4. Gray Level Run Length Matrix

La matriz de Secuencia de Longitud de Niveles de gris o GLRLM por sus siglas en inglés, es una matriz de la cual se es posible extraer ciertas características que permiten el análisis de textura en imágenes. La secuencia de longitud (run length) se entiende como un patrón de píxeles con cierta intensidad de gris en una determinada dirección con respecto a un píxel de referencia.

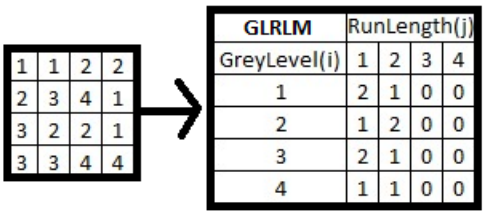


Figura 3.6: Ejemplo de la conformación de una GLRLM [Fuente: Elaboración propia]

La GLRLM es una matriz que se construye a partir de la ecuación 3.43.

$$GLRLM(x, y) = p(i, j|\theta) \tag{3.43}$$

Donde cada elemento de la matriz es la cantidad j de píxeles consecutivos con intensidad i en la dirección θ. Esto se puede apreciar mejor en la figura 3.6, donde se crea la matriz GLRLM correspondiente a un ángulo de 0°.

El método de GLRLM es una forma efectiva para lograr extraer las características estadísticas de textura de alto orden, algunas de ellas son:

- Run Emphasis (LRE)
- Gray level non-uniformity (GLN)
- Run length non-uniformity (RLN)
- Run Percentage (RP)

- Low Gray Level Run Emphasis (LGLRE)
- High Gray Level Run Emphasis (HGLRE).

3.3.4.3 Características basadas en descriptores espacio-frecuenciales

3.3.4.3.1. segmentación basada en el análisis de textura en fractales

La SFTA se presenta por primera vez en [Costa et al., 2012]. El algoritmo de extracción de características consiste en la descomposición de la imagen de entrada en un conjunto de imágenes binarias de las cuales se calculan las dimensiones fractales de los bordes de las regiones segmentadas, así se describen de manera óptima los patrones de textura de la imagen.

3.3.5 Selección de características

Con el fin de mejorar el desempeño de los clasificadores, es común que la extracción de características sea precedida por la etapa de selección de características, en la cual se implementan distintas técnicas para filtrar del vector original las que contengan menos información discriminante de las clases.

Para realizar esta selección existen distintas técnicas desarrolladas a partir de pruebas estadísticas.

3.3.5.1 Algoritmo de la mínima redundancia, máxima relevancia

El algoritmo MRMR encuentra un conjunto óptimo de características que son mutuamente y máximamente diferentes y pueden representar la variable de respuesta de manera efectiva [4]. Este se basa en la información mutua de las características.

La información mutua entre dos variables mide cuánta incertidumbre de una variable

[4] https://la.mathworks.com/help/stats/fscmrmr.html

puede reducirse conociendo la otra variable y está dada por la ecuación 3.44.

$$I(X,Z) = \Sigma_{i,j} P(X = x_i, Z = z_j) \cdot log \frac{P(X = x_i, Z = z_j)}{P(X = x_i)P(Z = z_j)} \tag{3.44}$$

Donde X y Z son variables discretas aleatorias y si son independientes, entonces I es igual a 0. Si X y Z son la misma variable aleatoria, entonces I es igual a la entropía de X.

El objetivo del algoritmo MRMR es encontrar un conjunto óptimo S de características que maximice V_S, la relevancia de S con respecto a una variable de respuesta y, y minimice W_S, la redundancia de S, donde V_S y W_S se definen con la información mutua I.

$$V_S = \frac{1}{|S|} \Sigma_{x \epsilon S} I(x, y) \tag{3.45}$$

$$W_S = \frac{1}{|S|^2} \Sigma_{x,z \epsilon S} I(x, z) \tag{3.46}$$

Donde $|S|$ es el número de características.

Luego, utilizando el cociente $\frac{V_x}{W_x}$, se le asigna un puntaje a cada característica x y se hace un ranking para determinar las más importantes según este algoritmo.

3.3.5.2 Características Univariante mediante pruebas de chi-cuadrado

Este algoritmo examina si cada variable predictora es independiente de una variable de respuesta mediante el uso de pruebas individuales de chi-cuadrado. Un valor pequeño de p indica que la variable predictora correspondiente depende de la variable de respuesta y, por lo tanto, es una característica importante[5].

[5]https://la.mathworks.com/help/stats/fscchi2.html

El puntaje en este algoritmo se calcula utilizando la ecuación 3.47. A medida que el valor de p sea más pequeño, la característica será más importante en el ranking de este algoritmo.

$$X^2_{score} = -log(p) \tag{3.47}$$

3.3.5.3 Selección de características mediante el análisis de componentes de vecindad para la clasificación

Esta técnica se basa en el análisis de componentes de la vecindad (NCA), el cual es un método de aprendizaje supervisado para la clasificación multivariable, similar al de k-vecinos cercanos. Al entrenar el clasificador, se generan los pesos para cada característica, los que corresponden a las características importantes contarán con pesos más altos mientras que las más irrelevantes tendrán valores cercanos a ceros y así podrán ser descartadas[6].

3.3.5.4 Rank importance of predictors using ReliefF or RReliefF algorithm

3.3.5.4.1. ReliefF

Este algoritmo encuentra los pesos de los predictores en el caso en que y es una variable categórica multiclase utilizando k-vecinos cercanos. El algoritmo penaliza los predictores que dan valores diferentes a los vecinos de la misma clase y recompensa los predictores que dan valores diferentes a los vecinos de diferentes clases.

ReliefF primero establece todos los pesos de los predictores W_j en 0. Luego, el algoritmo selecciona iterativamente una observación aleatoria x_r, encuentra las k observaciones más cercanas a x_r para cada clase y actualiza, para cada vecino más cercano x_q, todos los pesos para los predictores F_j siguiendo las ecuaciones 3.48 y 3.49 [7].

[6] https://la.mathworks.com/help/stats/fscnca.html
[7] https://la.mathworks.com/help/stats/relieff.html

Si x_r y x_q están en la misma clase:

$$W_j^i = W_j^{i-1} - \frac{\Delta j(x_r, x_q)}{m} drq \tag{3.48}$$

Si x_r y x_q están en distinta clase

$$W_j^i = W_j^{i-1} + \frac{p_{y_q}}{1 - p_{yr}} \cdot \frac{\Delta j(x_r, x_q)}{m} drq \tag{3.49}$$

Donde:

- W_i^i es el peso del predictor F_j en la iésima iteración.
- p_{y_r} es la probabilidad a priori de la clase a la cual pertenece x_r, y p_{y_q} es la probabilidad a priori de la clase a la cual pertenece x_q
- m es el numero de iteraciones
- $\Delta j(x_r, x_q)$ es la diferencia entre el valor del predictor F_j entre las observaciones x_r y x_q.
- d_{rq} es la función de distancia

3.3.5.4.2. RReliefF

Este algoritmo funciona de manera similar al ReliefF, sin embargo, este usa los pesos intermedios para computar los pesos finales del predictor.

Dados 2 vecinos cercanos, se asume que:

- W_{dy} es el peso de tener diferentes valores para la respuesta y.
- W_{dj} es el peso de tener diferentes valores para el predictor F_j.

- $W_{dy \wedge dj}$ es el peso de tener diferentes valores de respuesta y diferentes calores para el predictor.

El algoritmo primero establece los pesos en 0, luego iterativamente selecciona observaciones aleatorias x_r, encuentra las k observaciones cercanas y actualiza todos los pesos intermedios para cada vecino cercano x_q siguiendo las ecuaciones:

$$W_{dy}^{i} = W_{dy}^{i-1} + \Delta y(x_r, x_q) \cdot d_{rq} \tag{3.50}$$

$$W_{dj}^{i} = W_{dj}^{i-1} + \Delta j(x_r, x_q) \cdot d_{rq} \tag{3.51}$$

$$W_{dy \wedge dj}^{i} = W_{dy \wedge dj}^{i-1} + \Delta y(x_r, x_q) \cdot \Delta j(x_r, x_q) \cdot d_{rq} \tag{3.52}$$

Por último, calcula los pesos W_j (ecuación 3.53) después de actualizar por completo los pesos intermedios.

$$W_j = \frac{W_{dy \wedge dj}}{W_{dy}} - \frac{W_{dj} - W_{dy \wedge dj}}{m - W_{dy}} \tag{3.53}$$

3.3.6 Clasificadores basados en aprendizaje automático

3.3.6.1 Redes neuronales artificiales

Una RNA es un esquema de computación inspirado en la estructura que conforman las neuronas dentro del sistema nervioso de los seres humanos. Los modelos de RNA son capaces de encontrar relaciones o patrones de forma inductiva por medio de los algoritmos de aprendizaje basados en los datos existentes [Salas, 2004]. Una red neuronal combina diversas capas de procesamiento y utiliza elementos simples que operan en paralelo. Consta de una capa de entrada, una o varias capas ocultas y una capa de salida. Las capas están interconectadas mediante nodos, o neuronas; cada capa utiliza

la salida de la capa anterior como entrada[8].

Las redes neuronales supervisadas son entrenadas a partir de una serie de datos junto con su respectiva respuesta, por lo que resultan idóneas para modelar y controlar sistemas dinámicos, clasificar datos con ruido y predecir eventos futuros. Algunos tipos de redes son: feedforward, de base radial y dinámicas.

3.3.6.2 Máquinas de soporte vectorial

El algoritmo de clasificación binaria SVM busca un hiperplano óptimo que separe los datos en dos clases. Para las clases separables, el hiperplano óptimo maximiza un margen (espacio que no contiene ninguna observación) que le rodea, lo que crea límites para las clases positivas y negativas. Para las clases inseparables, el objetivo es el mismo, pero el algoritmo impone una penalización en la longitud del margen para cada observación que se encuentra en el lado equivocado de su límite de clase [9].

La función de puntuación lineal de la SVM es:

$$f(x) = x'\beta + b \tag{3.54}$$

Donde x es una observación o muestra, β contiene los coeficientes que definen un vector ortogonal al hiperplano y b es el bias.

3.3.6.3 K - vecinos cercanos

Asigna un vector de entrada a la clase predominante de los los k vecinos más próximos. Dependiendo de la cantidad de vecinos que pertenezcan a una clase, realizará la clasificación del dato nuevo. El algoritmo predice la clasificación de un dato x nuevo basándose en lo siguiente:

- Encuentra los puntos k dentro del dataset de entrenamiento más cercanos a x.

[8]https://la.mathworks.com/discovery/neural-network.html
[9]https://la.mathworks.com/help/stats/fitcsvm.html

- Encuentra la probabilidad y de que x pertenece a determinada clase para esos puntos más cercanos.
- Asigna la etiqueta de clasificación pertinente al que tenga la mayor probabilidad a posteriori entre los valores de y.

3.3.6.4 Modelo de clasificación bayesiana

Naive Bayes es un algoritmo de clasificación que aplica la estimación de la densidad de probabilidad a los datos.

El algoritmo aprovecha el teorema de Bayes y asume (ingenuamente, de ahí el nombre, *naive*) que los predictores son condicionalmente independientes, dada la clase. Aunque en la práctica se suele incumplir este supuesto, los clasificadores Bayes ingenuos tienden a producir distribuciones posteriores que son robustas a las estimaciones de densidad de clase sesgadas.

Los clasificadores Naive Bayes asignan las observaciones a la clase más probable (en otras palabras, la regla de decisión máxima a posteriori). Explícitamente, el algoritmo sigue estos pasos:

- Estima las densidades de probabilidad a los predictores dentro de cada clase con la ecuación:

$$p(x) = \frac{1}{(2\pi)^{d/2} \cdot |\Sigma|^{1/d}} \cdot e^{-\frac{1}{2}(x-\mu)^t \cdot \Sigma^{-1} \cdot (x-\mu)} \tag{3.55}$$

  Donde d es el número de características, μ es el vector de promedios y Σ es la matriz de covarianza.
- Se calcula la probabilidad a posteriori para cada clase.
- Se asigna la nueva muestra a la clase que tenga la mayor probabilidad a posteriori (Regla de decisión de Bayes).

3.3.7 Métricas para el desempeño de clasificadores

3.3.7.1 ANOVA

El Análisis de la varianza (ANOVA) es una fórmula estadística que se utiliza para comparar las varianzas entre las medias (o el promedio) de diferentes grupos de datos. Este es un procedimiento para determinar si la variación en la variable de respuesta deriva de la interacción de los distintos grupos de población o de los propios grupos. Uno de los datos que se puede extraer del ANOVA es el valor p o nivel de significancia.

3.3.7.1.1. Valor p

El nivel de significancia o valor p, es la probabilidad de rechazar la hipótesis nula cuando es verdadera. Es una medida de la probabilidad de que la diferencia de resultado se deba al azar.

4. Metodología propuesta para la detección de microcalcificaciones en mamografías digitales

En este capítulo se presenta la metodología usada para el desarrollo de un sistema de detección automática de microcalcificaciones MCs mediante el análisis de mamografías digitales estructurado en siete etapas: (1). selección de la base de datos, (2). preprocesamiento, (3). constitución de la base de datos para el entrenamiento de las máquinas de aprendizaje, (4). extracción de características, (5). selección de características, (6). entrenamiento de máquinas de clasificación, y (7). inferencia y postvalidación.

En la etapa de selección de la base de datos, se muestra la evaluación para la escogencia del conjunto de datos en función del estado del arte reportado para trabajos similares y las características relevantes de la base de datos. En la etapa de preprocesamiento, se aplicó a las mamografías digitales seleccionadas técnicas de visión artificial para la mejora de contraste, supresión de artefactos y eliminación del tejido pectoral. En la etapa de la constitución de la base de datos para el entrenamiento de las máquinas de aprendizaje se plantea la estrategia utilizada para la creación de la base de datos que se utilizará en este proyecto. Posteriormente, se implementó un proceso de aumento de datos sobre la base de datos utilizada con el fin de tener un mayor número de muestras para las etapas posteriores. Luego, en la etapa de extracción de características, usando la base de datos aumentada, se extrajeron 149 características basadas en descriptores estadísticos, de textura y espacio-frecuenciales para examinar una amplia gama de técnicas con el fin de robustecer el conjunto de características que se obtendrá; con lo anterior, en la etapa de selección de características se aplicaron técnicas para realizar una ponderación de las mismas y así reducir la dimensionalidad del problema. En la etapa de entrenamiento, se probaron cuatro clasificadores: Red neuronal Artificial (RNA) tipo MLP, Máquinas de vectores de soporte (SVM), K vecinos más cercanos (K-NN) y el clasificador Bayesiano ingenuo (NB). Finalmente, en la etapa de inferencia y postvalidación se presenta la estrategia de evaluación seleccionada para el sistema.

A continuación, se describen cada una de las etapas mostrando en detalle el desarrollo de las mismas.

4.1 Selección de la base de datos

En el desarrollo y estudio de sistemas de visión artificial son utilizadas múltiples bases de datos de mamografías digitales, pero, la mayoría de bases de datos presentadas en la literatura son de uso restringido, dado que normalmente son construidas entre grupos de investigación y entidades médicas como clínicas y hospitales de la región a la cual pertenezca el grupo de investigación.

Así las cosas, existen solo algunas bases de datos de mamografías digitales de libre uso que han sido utilizadas a lo largo del tiempo, reportando un alto índice de confiabilidad como se afirma en [Moreira et al., 2012]. A continuación se describen las cuatro bases de datos de libre acceso más utilizadas:

(1). MIAS es la base de datos más antigua, cuenta con una población de 161 casos, todas sus mamografías fueron tomadas en la vista tipo Medio Lateral Oblicua (MLO), posee una diversidad de tamaños en las imágenes, así como diversidad en tipos de densidades mamarias, una resolución espacial de 5 μm por píxel y las anotaciones realizadas por los expertos consisten en coordenadas (x,y) sobre la imagen, asociadas a un radio dado que representa un área circular con presencia de la región afectada.

(2). MiniMIAS es el resultado del reescalamiento en tamaño de la base de datos MIAS, por lo cual todas las imágenes cuentan con un tamaño uniforme de 1024x1024 píxeles lo que forzó a que su resolución espacial se redujera de 50 μm por píxel a 200 μm por píxel. El motivo de este reescalamiento surgió con el objetivo de estandarizar el tamaño de las imágenes para trabajos que no tuvieran inconvenientes en la reducción espacial realizada.

(3). DDSM cuenta con una población de 2620 casos, todas sus mamografías fueron tomadas en la vista Medio Lateral Oblicua (MLO) y Cranio-Caudal (CC), posee imágenes de tamaño fijo y diversidad en tipos de densidades mamarias. Las anotaciones realizadas por los expertos demarcan el borde de los píxeles de la región

afectada. No cuenta con información sobre la resolución espacial de sus imágenes.

(4). BancoWeb cuenta con una población de 320 casos, todas su mamografías fueron tomadas en la vista Medio Lateral Oblicua (MLO) y Cráneo-Caudal (CC), posee imágenes de tamaño fijo y diversidad en tipos de densidades mamarias. Las anotaciones realizadas por los expertos demarcan las coordenadas de un rectángulo del área afectada, sin embargo, no todas las imágenes cuentan con ésta información. No cuenta con información sobre la resolución espacial de sus imágenes.

Dentro de las características más importantes observadas en el resumen de la tabla 4.1, se encuentra la información de la resolución espacial de las imágenes, dado que la patología microcalcificaciones MCs reporta unas dimensiones típicas que rondan entre los 20 μm y 100 μm como se menciona en [Henrot et al., 2014] y como se puede observar las bases de datos DDSM y BancoWeb no cuentan con esa información, por tanto fueron descartadas para su uso en este trabajo. Por lo anterior, MIAS y MiniMIAS presentaron un especial interés para su escogencia en el desarrollo de este trabajo, donde finalmente se decantó el uso de MIAS como la base de datos seleccionada gracias a que la resolución espacial de las mamografías se encuentra dentro del rango típico de un tamaño posible de microcalcificaciones MCs, resaltando que este trabajo pretendió especializarse en la detección de regiones con este tipo de patologías y no de otras más. Dicho tamaño de resolución espacial cuenta además con un interés particular, ya que como se verá más adelante, es un factor importante para la escogencia del área de interés en la creación de la base datos que contendrá la región con presencia de MCs.

Por otra parte, MIAS cuenta con un tamaño de las imágenes así: a. **pequeña (s)**: 1600x4320 píxeles, b. **mediana (m)**: 2048x4320 píxeles, c. **larga (l)**: 2600x4320 píxeles y d. **extra-larga (xl)**: 5200x4320 píxeles, factor que no influye en el desarrollo del trabajo, puesto como se verá más adelante, mediante el método de ventanas deslizantes para la clasificación, no importa el tamaño de la mamografía inspeccionada.

La base de datos MIAS cuenta con siete tipos de lesiones presentadas en las mamografías, de las cuales 25 imágenes que contienen contienen casos de de microcalcificaciones MCs, las cuales fueron seleccionadas como base inicial para el desarrollo de este trabajo.

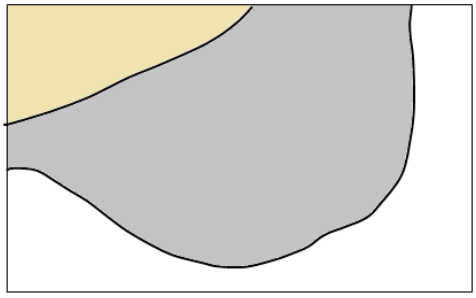
La tabla 4.1, resume la información anteriormente presentada.

	(1). MIAS	(2). MiniMIAS	(3). DDSM	(4). BancoWeb
Año	1994	1995	1999	2010
Número de casos	161	161	2620	320
Número de imágenes	322	322	10480	1400
Tipo de vista	MLO	MLO	MLO y CC	MLO y CC
Tipo de lesiones	Todas	Todas	Todas	Todas
Tipos de densidades mamarias	SI	SI	SI	SI
Tipo de imagen	PGM	PGM	LJPEG	TIFF
Resolución	8 bits/píxel	8 bits/píxel	8 o 16 bits/píxel	12 bits/píxel
Resolución espacial	50 micrómetros/píxel	200 micrómetros/píxel	-	-
Anotaciones (Ground truth)	Coordenadas (x,y) y radio de círculo del área afectada	Coordenadas (x,y) y radio de círculo del área afectada	Borde píxeles del área afectada	Coordenadas rectángulo del área afectada (No todas las mamografías)

Tabla 4.1: Comparación de las bases de datos utilizadas en la literatura. Fuente: elaboración propia con datos tomados de [Moreira et al., 2012].

Adicionalmente, un aspecto importante a tener en cuenta dentro de las base de datos de mamografías es la disposición espacial de la mama en una mamografía digital dependiendo si es un seno izquierdo o uno derecho.

(a) Disposición de una mamografía correspondiente a un seno derecho.



(b) Disposición de una mamografía correspondiente a un seno izquierdo.

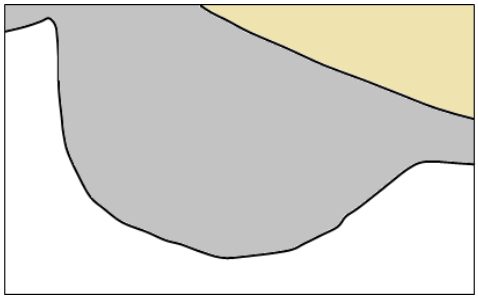


Figura 4.1: Orientación de la mamografía según el tipo de seno en la base de datos MIAS. [Fuente Propia].

Las Figuras 4.1a y 4.1b representan la orientación de un seno derecho e izquierdo respectivamente. Se ha rellenado el tejido pectoral de color amarillo, el resto del seno de color gris claro (zona donde normalmente se encuentran las MCs) y el fondo de la mamografía de color blanco.

Según lo mencionado anteriormente y en aras de desarrollar un algoritmo que realice la segmentación automática de la zona de interés (región gris claro en la figura

4.1), es importante normalizar la orientación de las mamografías y así facilitar etapas subsecuentes, proceso que se mostrará en detalle más adelante.

Por otro lado, es relevante a mencionar que dentro de cada imagen mamográfica se pueden encontrar elementos como artefactos y ruido, además, dentro del grupo de 25 imágenes seleccionadas, el 48 % de las mamografías tiene presencia de artefactos y el 100 % de ruido.

- **Artefactos**: En los métodos tradicionales de rayos X, es común encontrar las etiquetas de identificación del paciente y número de mamografía, las cuales eran impresas sobre el papel radiológico inclusive antes de generar la exposición del paciente en la máquina de rayos X.

- **Ruido**: La calidad de la mamografía convertida en imagen digital es dependiente del tipo de equipo digitalizador, y es común que partículas de polvo, apertura de luz, problemas mecánicos y otros estén presentes en el proceso de digitalización causando un efecto de ruido digital.

La figura 4.2 presenta un ejemplo real perteneciente a las 25 mamografías mostrando los elementos mencionados anteriormente, así como la presencia del tejido mamario y el tejido pectoral. Con el fin de de resaltar las partes comúnmente encontradas sobre la imagen mamográfica (además del tejido pectoral y la mama), se realizaron dibujos sobre dichos objetos y se encuentran etiquetados en la figura 4.2.

Adicionalmente, cabe resaltar que en la figura 4.2, en la zona izquierda se presenta un recuadro de aumento que corresponde a la zona en donde los expertos, creadores de la base de datos, especificaron una zona con presencia de MCs. Así mismo, se aprecia el grado de dificultad que presenta la evaluación de este tipo de anormalidades (puntos blancos brillantes) dentro de una mamografía digital debido a su diminuto tamaño, incluso después del zoom hecho sobre la microcalcificación individual. Además, como se puede observar en la figura 4.2, el seno y el fondo no están bien contrastado y, por ende, el borde del seno no está claramente definido.

Debido a la presencia de fondo negro, los artefactos (más claros que el fondo) pueden llegar a incidir en el momento de la segmentación si no se tratan de manera adecuada. Por esta razón, se necesitará una etapa de preprocesamiento para remover los elementos que puedan interferir y se presentará en detalle en etapas posteriores.

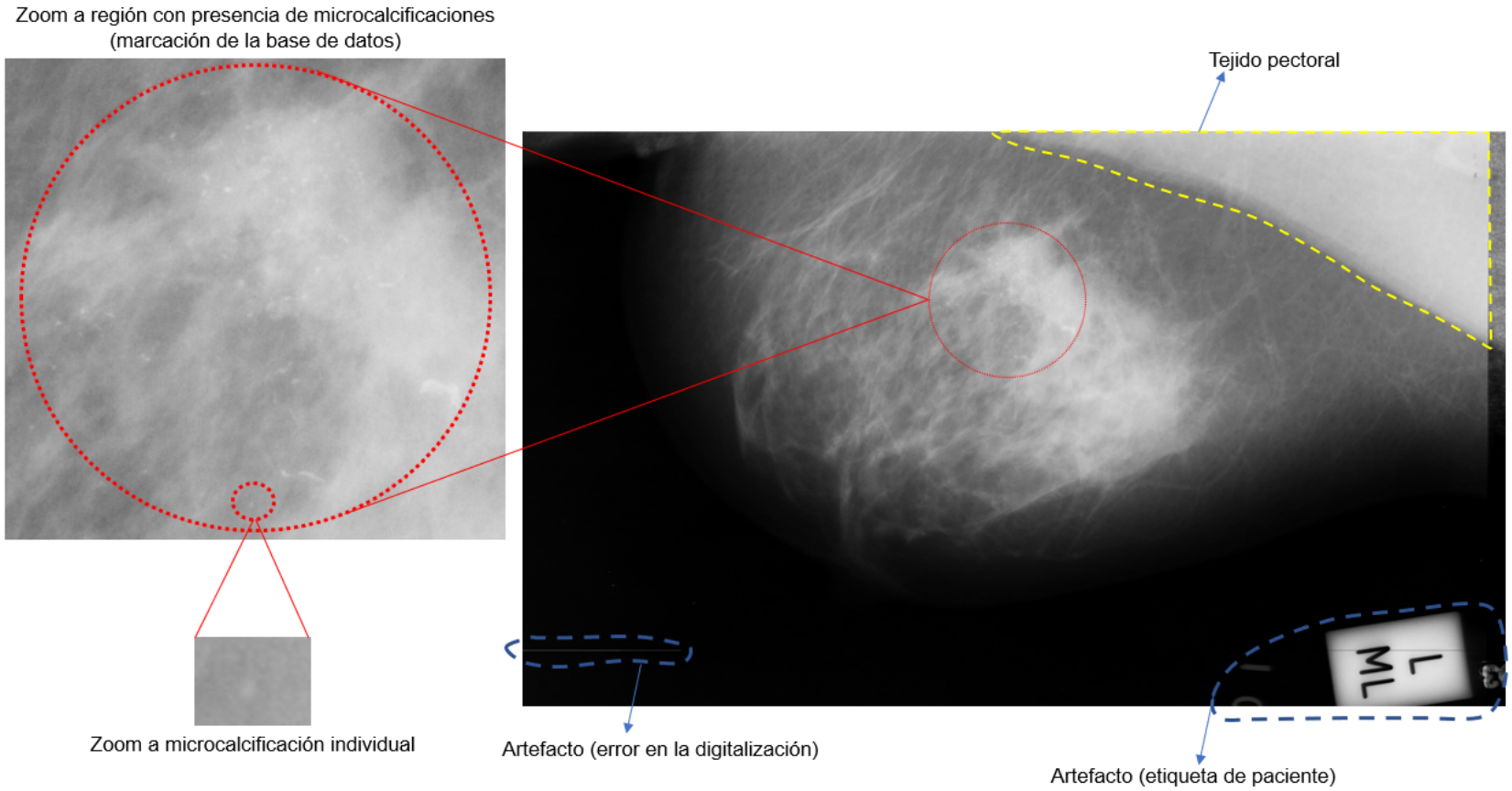


Figura 4.2: Mamografía de seno derecho con anotaciones de ruido, artefactos, tejido pectoral y presencia de MCs. Imagen N° 209 base de datos MIAS. [Fuente Propia].

4.2 Preprocesamiento

Esta fase tiene como objetivo principal estandarizar la orientación espacial de todas las mamografías, mejorar el contraste en la imagen para delimitar el borde entre el seno y el fondo, identificar el tejido pectoral para después ser eliminado y facilitar la eliminación de los artefactos presentes en las imágenes. A continuación se detalla el desarrollo de estos procesos.

4.2.1 Estandarización de la orientación espacial

Como se mencionó anteriormente, las mamografías poseen una orientación espacial específica dependiente de si el seno capturado es derecho o izquierdo. Por tanto, es necesario realizar un procedimiento para mantener una única orientación espacial de la imagen con el fin de tener un único algoritmo que no dependa de la orientación de la imagen.

Analizando el histograma de una imagen mamográfica, se puede apreciar que la característica predominante en una mamografía: la presencia del fondo negro permite observar una distribución elevada de píxeles hacia los niveles de gris de menor magnitud presentes en el histograma. Por tanto, es evidente que las demás partes de la imagen son: el seno, los artefactos y el tejido pectoral se encuentran en su mayoría distribuidos hacía los niveles de gris de magnitud superior. Lo anterior se puede apreciar en la figura 4.3. Adicionalmente, como se puede observar en el histograma que a parte de haber una concentración de niveles de gris negros (recuadro rojo izquierdo en la figura 4.3) en la imagen, el tejido pectoral alberga una homogeneidad en sus niveles de gris donde particularmente se encuentran distribuidos hacia las magnitudes más altas en el histograma (tendencia al color blanco, recuadro rojo derecho en la figura 4.3).

Ahora, con el fin de determinar la orientación espacial del seno, se parte del hecho de que en todas las imágenes (sea seno derecho o izquierdo) existe la presencia del tejido pectoral ubicado en la parte superior. Debido a esta particularidad, se decide dividir la imagen en cuatro ventanas iguales, de las cuales se decide analizar el histograma de las regiones superiores de manera independiente. Las figuras 4.4a y 4.4b representan la división mencionada.

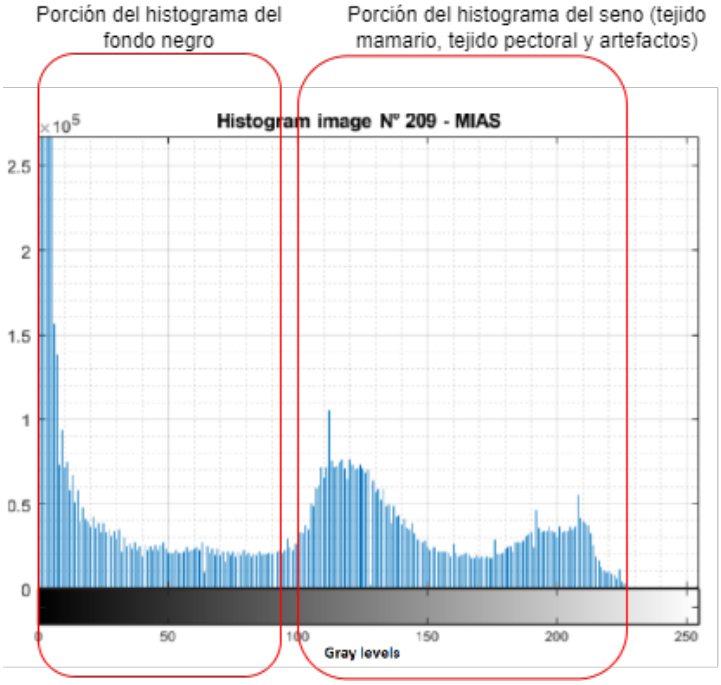
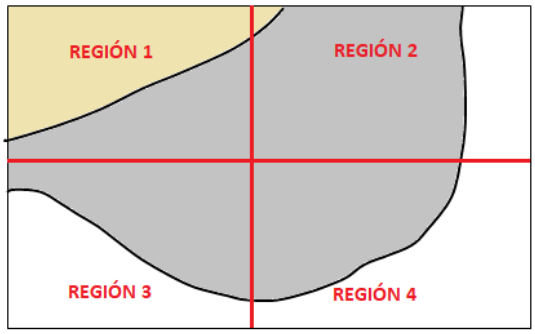


Figura 4.3: Histograma de Imagen N° 209 base de datos MIAS.

Con el análisis independiente de las regiones superiores, se aprecia que donde se encuentra mayoritariamente el tejido pectoral *(Región 2 - seno izquierdo y Región 1 - seno derecho)* en su histograma predomina una distribución de niveles de grises con magnitudes más altas cercanas al blanco como se visualiza en la figura 4.5.

(a) División en cuatro regiones seno derecho.



(b) División en cuatro regiones seno izquierdo.

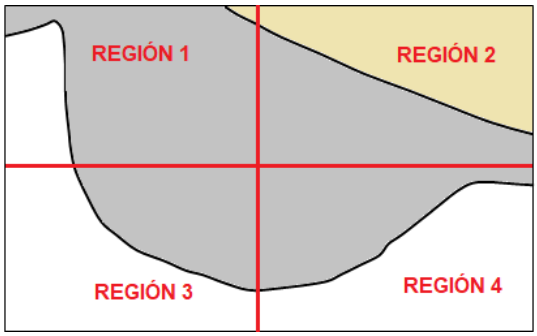


Figura 4.4: División del seno en cuatro regiones. [Fuente Propia].

Teniendo en cuenta el análisis visual realizado, la zona pectoral es más brillante que la parte del seno, debido a esto, se decide extraer la característica de brillo (calculando

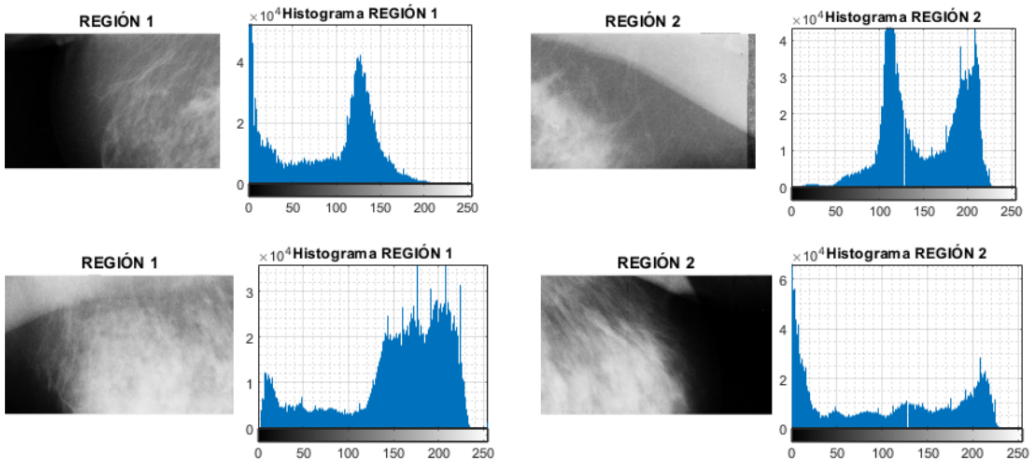


Figura 4.5: Histogramas de las regiones 1 y 2 para mamografías de seno izquierdo (parte superior) y de seno derecho (parte inferior). Fuente propia.

el promedio de los valores de gris del histograma) sobre cada región independiente y verificar la repetibilidad en las regiones del tejido pectoral para todas las imágenes. Teniendo la imagen $I(x, y)$ de dimensión n x m, se utiliza la ecuación 4.1 para el cálculo de las 25 imágenes.

$$Brillo = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} I(i,j)}{n \cdot m} \tag{4.1}$$

Las figuras 4.6a y 4.6b son el diagrama de cajas del cálculo del brillo de las mamografías de senos izquierdos y derechos respectivamente para cada región, incluyendo las *Regiones 3* y *4* con el objetivo de verificar que su brillo corresponde a una magnitud más baja que las *Regiones 1* y *2*.Así mismo, se puede verificar como en ambos casos *(Región 2 - senos izquierdo y Región 1 - senos derechos)*, los valores de la mediana (línea roja dentro de los recuadros azules en las figuras 4.6a y 4.6b) del brillo son los más altos, diferenciándose de las demás regiones.

Una vez comprobado que la región con presencia de tejido pectoral contiene el brillo más alto de todas las regiones para la mayoría de las imágenes, se diseña un algoritmo para rotar de manera automática la imagen de entrada y posicionarla espacialmente de forma que el tejido pectoral siempre quede ubicado en la esquina superior derecha de la imagen. Así, para imágenes de seno derecho, se aplica una rotación sencilla de 45°,

(a) Promedios por cada región de senos izquierdos

(b) Promedios por cada región de senos derechos

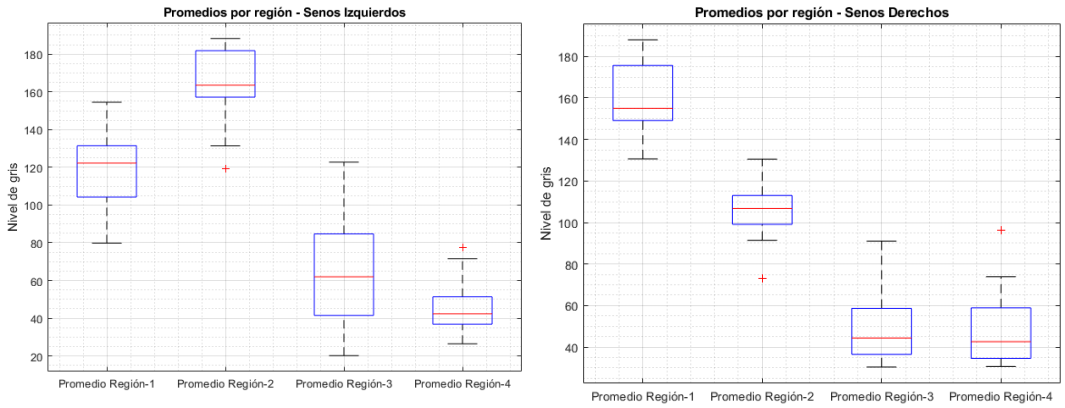


Figura 4.6: Diagrama de cajas de la distribución por cada región 1, 2, 3 y 4 para senos izquierdos y derechos.

mientras que para el seno izquierdo se aplica una rotación de -45° y posteriormente una operación de *flip*, estos pasos se ilustran en la figura 4.7. Es importante resaltar que la disposición espacial de la imagen como se sugiere en la figura 4.7 (imagen con pectoral en la parte superior derecha) se ha seleccionado de manera aleatoria y no obedece a ningún criterio específico.

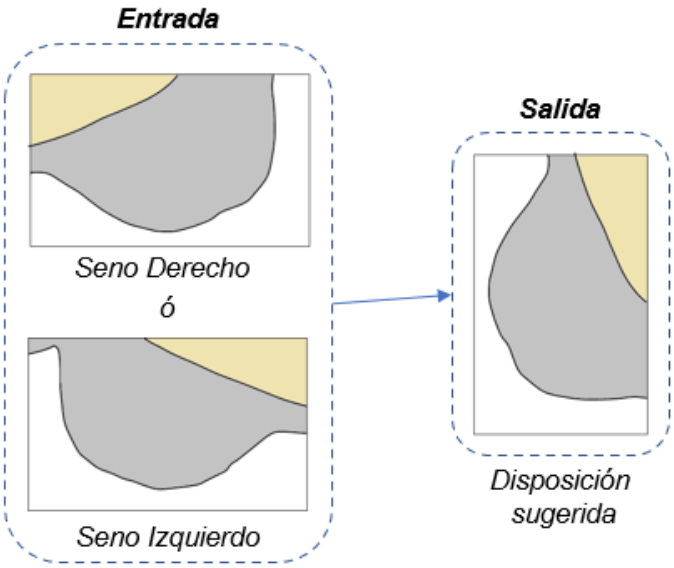


Figura 4.7: Disposición espacial deseada para la mamografía. [Fuente Propia].

Esta estrategia funcionó apropiadamente para 24 de las 25 imágenes del dataset MIAS en el grupo de calcificaciones, esto se debe a que esta imagen presenta un tejido mamario extremadamente denso y atípico, impidiendo que el algoritmo diseñado rotara la imagen de manera correcta, por esta razón, la rotación se hizo de manera manual.

4.2.2 Mejora del contraste

Una vez obtenido un protocolo de rotación automático de imagen presentado en 4.2.1, el siguiente objetivo consiste en resaltar las características relevantes en la imagen para facilitar la delimitación el seno y el tejido pectoral. Como se había mencionado anteriormente en 2.4.1, una tarea desafiante que conlleva el uso de imágenes mamográficas y en especial las de tipo SFM, es el bajo contraste que presenta la imagen, dado que no permite diferenciar de manera clara los elementos de interés para esta fase: bordes.

Como se pudo verificar anteriormente en la imagen 4.2, la delimitación del borde del seno y el fondo es muy suave. Para lograr una mejora de contraste en la imagen, es necesario realizar una modificación sobre el histograma teniendo en cuenta la siguiente premisa: ***mejorar la iluminación en las zonas más oscuras de la imagen dado que el borde del seno se encuentra en esta región***.

Si bien existen diferentes familias de funciones para comprimir o ampliar el rango dinámico del brillo de una imagen, la transformación de potencia o **corrección gama** fue seleccionada debido a que permite variar las zonas con poca iluminación[Kumari et al., 2014], incrementando el contraste y siendo de gran ayuda para resaltar el borde entre el seno y el fondo.

Como se mencionó en el capítulo 3, valores de la constante γ menores a 1 permiten elevar el nivel de gris de salida para píxeles de entrada con nivel de gris bajo, sin afectar demasiado los niveles de gris altos.

Teniendo la imagen $I(x, y)$ de dimensión n x m, se utiliza la ecuación 4.2 que define

la transformada gama de una imagen.

$$g(x, y) = c \cdot I(x, y)^{\gamma} \tag{4.2}$$

De la ecuación 4.2, el valor de c es la constante de escalamiento para obtener los valores de salida en un rango dinámico deseado y esta dado por 4.3.

$$c = \frac{255}{max(I(x, y)^{\gamma})} \tag{4.3}$$

Con el objetivo de realizar una evaluación visual rápida, se toma una imagen de muestra y se aplica la transformación gama evaluando un rango de valores de gama comprendido desde $\gamma = 0.1$ hasta 1.9 en intervalos de 0.1, los resultados se observan en la figura 4.8.

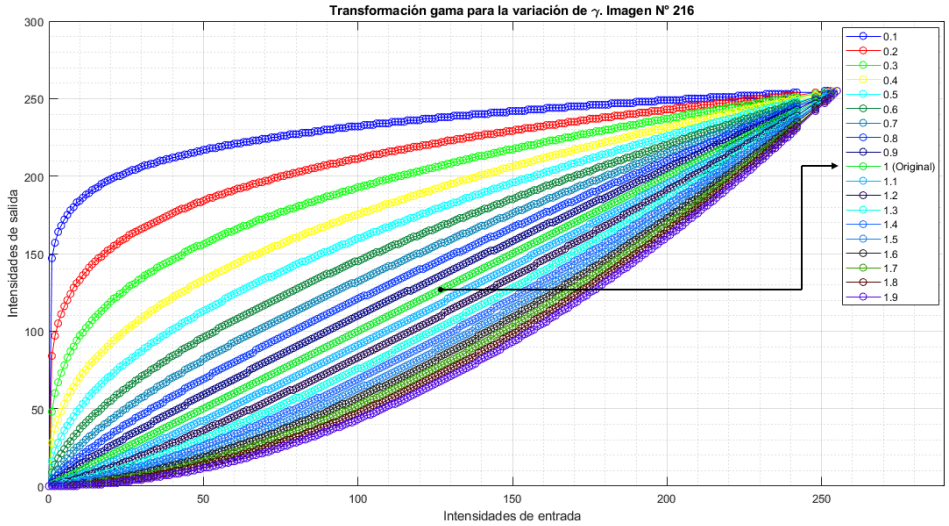


Figura 4.8: Intensidades de salida en función de las intensidades de entrada con variaciones del valor de γ. Imagen N° 216 base de datos MIAS.

En la figura 4.8 se puede observar las intensidades de salida en función de las intensidades de entrada para la imagen N° 216 de la base de datos MIAS.

La figura 4.9 presenta las imágenes resultantes para cada una de las transformaciones

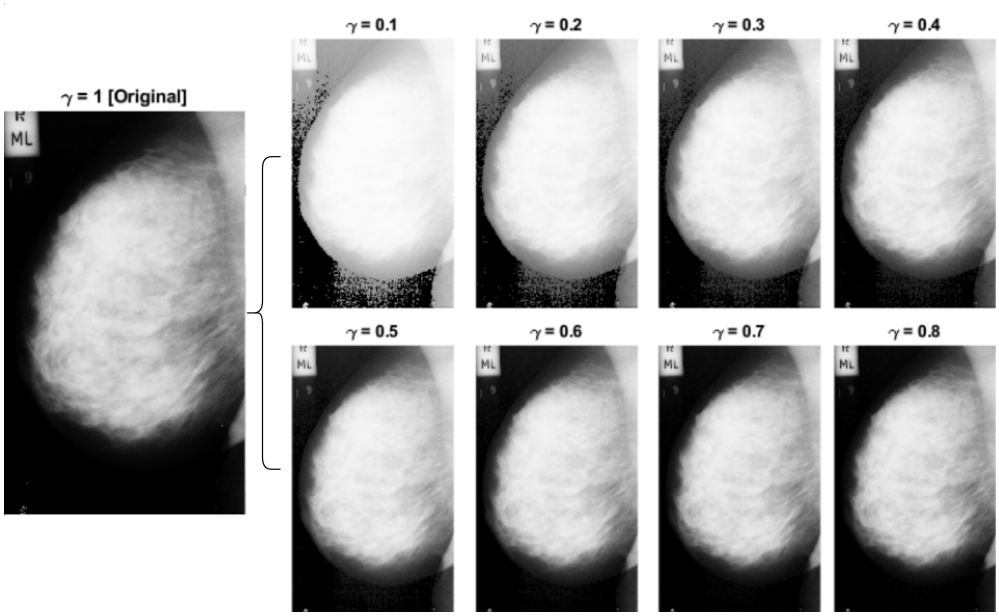


Figura 4.9: Imagen con transformada gama con valor $\gamma = 1$ (derecha) y transformadas gama para la Imagen N° 216 con variación de γ desde 0.1 hasta 0.8 (izquierda). Fuente propia.

y variación de γ. Como se esperaba, para valores del coeficiente gama γ menores a 1, las zonas que originalmente se encuentran con poca iluminación son resaltadas pudiéndose identificar de manera más eficiente el borde del seno. La figura 4.9 también muestra la imagen de salida siendo aplicada la transformación con coeficiente gama $\gamma = 1$, que permanece igual que la imagen original conservando la relación lineal entre intensidad de salida e intensidad de entrada como se verifica en la figura 4.8.

A pesar que, de manera visual se puede observar que la escogencia de un valor del coeficiente γ menor a 1 mejora el contraste de la imagen y permite resaltar el borde del seno, se utilizaron tres métricas comúnmente usadas para evaluar la calidad de una imagen presentadas en [Ma et al., 2019].

Como se mencionó en el capítulo 3, existen distintas métricas para evaluar la calidad de una imagen $I(x, y)$ de dimensión n x m. Ahora, aplicando las ecuaciones pertinentes para cada métrica a las imágenes obtenidas en 4.9, las figuras 4.10 4.11 y 4.12. presentan el valor de cada una de las métricas *STD, AG* y *SF* respectivamente en función de la variación del coeficiente gama γ para las 25 imágenes.

Se observa en la figura 4.10 que existe una división de tendencias en dos grupos de imágenes. El primero presenta valores elevados de STD para un coeficiente gama $\gamma = 0.1$ y va decrementando conforme gama va creciendo. El segundo grupo, nuevamente al igual que para la métrica SF, presenta un intervalo de valores de gama para el cual el valor de la métrica STD es el más alto. El intervalo en mención va aproximadamente desde $\gamma = 0.4$ hasta $\gamma = 0.8$.

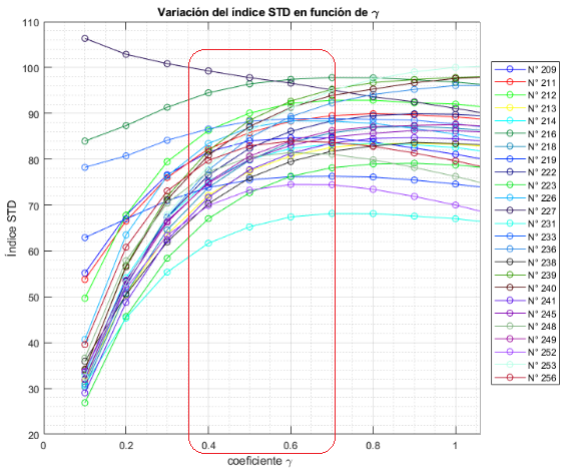


Figura 4.10: Métrica desviación estándar (STD) en función de la variación de la constante gama γ para las 25 imágenes.

La métrica AG muestra en la figura 4.11 un comportamiento donde se ve nuevamente una división en dos grupos.

Para el primero, las imágenes presentan el mayor valor de AG cuando el coeficiente gama γ es 0.1 y va decrementando conforme γ va creciendo. El segundo grupo presenta una tendencia inicialmente de magnitud baja en la métrica AG para valores bajos del coeficiente gama γ, un crecimiento considerable desde $\gamma = 0.4$ hasta $\gamma = 0.8$ y una tendencia de disminución constante para valores de gama superiores a $\gamma = 0.8$. Lo anterior indica que, el índice refleja una calidad de imagen contrastada prominentemente mayor en el intervalo de valores de coeficiente gama mencionado.

Desde otra perspectiva, la métrica SF muestra un comportamiento similar al de la métrica STD. Para el primero, las imágenes presentan el mayor valor de SF cuando el coeficiente gama γ es 0.1 y va decrementando conforme gama γ va creciendo hasta tender a estabilizarse. Por otro lado, el segundo grupo presenta una tendencia en el que

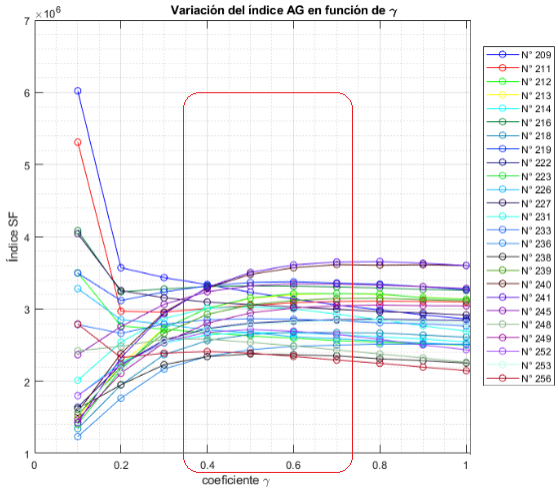


Figura 4.11: Métrica gradiente promedio (AG) en función de la variación de la constante gama γ para las 25 imágenes.

aproximadamente desde $\gamma = 0.1$ hasta $\gamma = 0.4$ existe un crecimiento de la métrica SF, sin embargo a partir de $\gamma = 0.4$, el valor de SF tiende a ser constante. De lo anterior, se puede concluir que evidentemente para valores pequeños (tendiendo al 0) del coeficiente gama, el contraste obtenido en la imagen es muy alto, por tanto, se observa que para ambos grupos el valor de SF es alto, indicando una riqueza en textura y bordes. Por otro lado, cabe resaltar que la presencia de ambos grupos puede ser causada a que algunas de las imágenes presentan un nivel de ruido y artefactos elevado a diferencia de otras que lo presentan de manera moderada o no presentan.

De la anterior visualización de métricas SF, AG y STD, se puede concluir que efectivamente para valores de coeficiente gama γ menores a 1 se mejora el contraste en general de la imagen. Debido a que existen dos grupos de imágenes que responden de manera diferente en cada una de las métricas, se decidió finalmente realizar un diagrama de cajas evaluando cada coeficiente gama γ para las 25 imágenes con el propósito de poder visualizar la distribución de las métricas de manera independiente para cada uno de los valores del coeficiente gama γ y así poder establecer una valor adecuado de coeficiente gama.

Gracias a la información de las métricas obtenidas anteriormente para el grupo de 25 imágenes se pude inferir que:

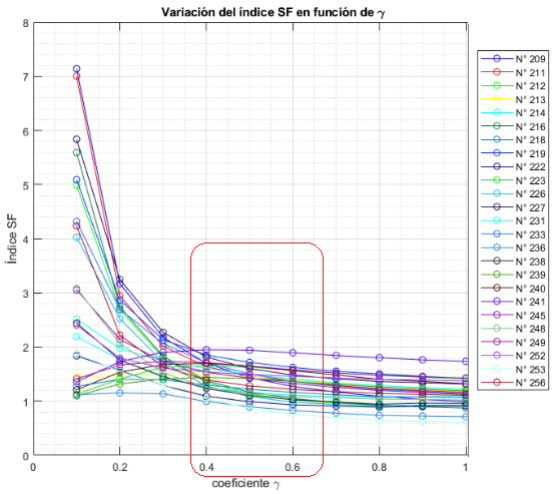


Figura 4.12: Métrica frecuencia espacial (SF) en función de la variación de la constante gama γ para las 25 imágenes.

- Existen dos grupos de imágenes identificables; en ambos casos los valores más altos obtenidos en las métricas utilizadas se dan para valores pequeños del coeficiente gamma ($\gamma < 0.8$).
- A pesar de que el 48 % de las imágenes presenta los más altos valores registrados de las métricas cuando el coeficiente gama es $\gamma = 0.1$, utilizar una imagen transformada con éste coeficiente resalta el borde del seno de una manera excesiva e incluso se potencia el contraste del ruido lo que afecta de manera directa el borde del seno como se puede visualizar en la figura 4.9.
- Dos de las tres métricas utilizadas (AG y STD), sugieren que en el intervalo del coeficiente gama $\gamma \geq 0.4$ y $\gamma \leq 0.8$ se presenta el mejor desempeño en cuanto contraste y riqueza en bordes se trata.

Así mismo, conforme el análisis realizado y lo mencionado anteriormente, se decidió seleccionar el valor del coeficiente gama a utilizar de manera cualitativa evaluando visualmente los resultados obtenidos en el intervalo del coeficiente gama $\gamma \geq 0.4$ y $\gamma \leq 0.8$.

Teniendo en cuenta la evaluación cualitativa, se considera que observando la imagen de muestra presentada en la figura 4.9, los resultados de las transformaciones para valores de $\gamma > 0.6$ no permiten distinguir de manera fácil el borde del seno, mientras

que para valores $\gamma \geq 0.4$ y $\gamma \leq 0.5$ si. Por tanto se decidió tomar un valor intermedio entre dicho intervalo, obteniendo así un valor de coeficiente $\gamma = 0.45$.

La figura 4.13 presenta el resultado para la imagen de muestra habiéndose aplicado la corrección gama con coeficiente $\gamma = 0.45$ y se puede verificar como se resalta el borde del seno sin potencializar las zonas con presencia de ruido.

(a) Imagen de muestra e imagen aplicada corrección gama.

(b) Intensidades de salida en función de las intensidades de entrada.

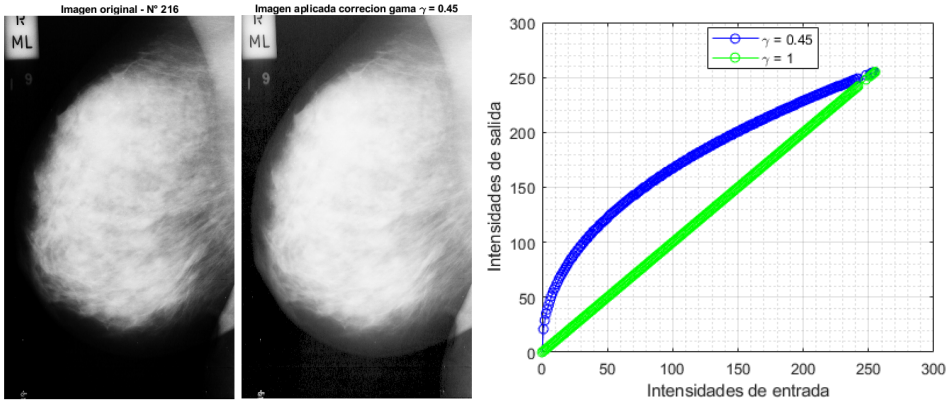


Figura 4.13: Ejemplo imagen N° 216 y corrección gama con el coeficiente γ seleccionado.

4.2.3 Delimitación del seno

Esta fase tiene como objetivo encontrar una máscara binaria que represente el seno y así suprimir los elementos presentes en el fondo de la imagen como artefactos, etiquetas y/o ruido. Por otro lado, también se pretende eliminar la región de imagen correspondiente al tejido pectoral.

Se realizará una primera etapa de segmentación, en la cual se separará el fondo del seno y el tejido pectoral. Posteriormente, en una segunda etapa se segmentará únicamente el seno, retirando el tejido pectoral.

4.2.3.1 Segmentación del seno y tejido pectoral

Una vez terminada la mejora del contraste para realzar el borde del seno, es posible realizar la segmentación del seno junto con el tejido pectoral, encontrando una máscara que permita separar esta zona del fondo. A pesar de que existen diferentes técnicas para realizar segmentación sobre imágenes, se decidió realizar este proceso mediante la técnica basada en el **Umbral por Histograma del Gradiente** mencionada en el capítulo 3.

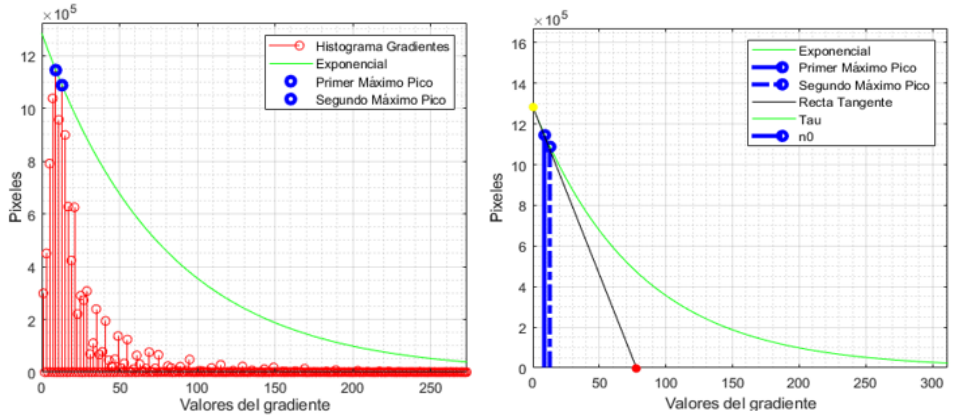


Figura 4.14: Histograma del gradiente de la imagen N° 248 (izquierda); envolvente y recta tangente del histograma del gradiente de la imagen N°248 (derecha).

La figura 4.14 muestra un ejemplo de aplicación de la técnica del histograma del gradiente, utilizando la imagen N° 248. Se puede apreciar en el lado derecho que el algoritmo de la técnica encuentra los dos valores picos del histograma. Posteriormente, en el lado izquierdo se utilizan dichos valores para calcular la recta tangente y encontrar el intercepto con el eje x (τ). El valor de τ encontrado para la imagen N° 248 fue de 77.85 y cabe resaltar que este varía en función de la imagen.

Es importante resaltar que el cálculo de la imagen del gradiente horizontal y vertical requerido por el algoritmo del histograma del gradiente se realizó utilizando el operador Sobel (que suele ser de los más utilizados en la detección de bordes), con kernel vertical y kernel horizontal mostrados en 4.4 y 4.5 y tomando como resultado los valores absolutos

de la operación.

$$Vertical = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{4.4}$$

$$Horizontal = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{4.5}$$

Por otro lado, la imagen de salida final, corresponde a un proceso de binarización utilizando como umbral el resultado τ obtenido por la metodología del histograma del gradiente. La binarización se basa en 4.6.

$$I_{binaria} = \begin{cases} 1 \rightarrow I_{gamma}(x,y) < \tau \\ 0 \rightarrow I_{gamma}(x,y) \geq \tau \end{cases} \tag{4.6}$$

Por lo anterior, se presenta en la la tabla 4.2 los valores de umbral τ obtenidos para todas las imágenes. Es importante resaltar que el valor de umbralización τ obtenido es decimal, por tanto, para realizar la evaluación de la ecuación 4.6 fue necesario redondear cada valor de τ al número entero más cercano mayor o igual que éste.

En la imagen 4.15 se presenta la visualización de los resultados de cada una de las imágenes binarias aplicando la ecuación 4.6 y usando los umbrales de la tabla 4.2. En la mayoría de los casos, las imágenes binarias representan adecuadamente el seno y se distingue el fondo, sin embargo, en algunas otras se visualizan regiones con presencia de artefactos como se observa en las demarcaciones de color rojo en la figura 4.15.

N° Imagen	209	211	212	213	214	216	218	219	222	223	226	227	231	233	236	238	239	240	241	245	248	249	252	253	256
Umbral	53.4	64.9	84.4	81.2	82.6	69.7	87.4	78.7	87.6	95.4	77.4	63.7	98.4	65.4	98.6	95.4	80.4	78.4	81.4	80.6	77.85	91.4	84.4	88.4	80.4

Tabla 4.2: Umbral obtenido por cada imagen con el método del histograma del gradiente.

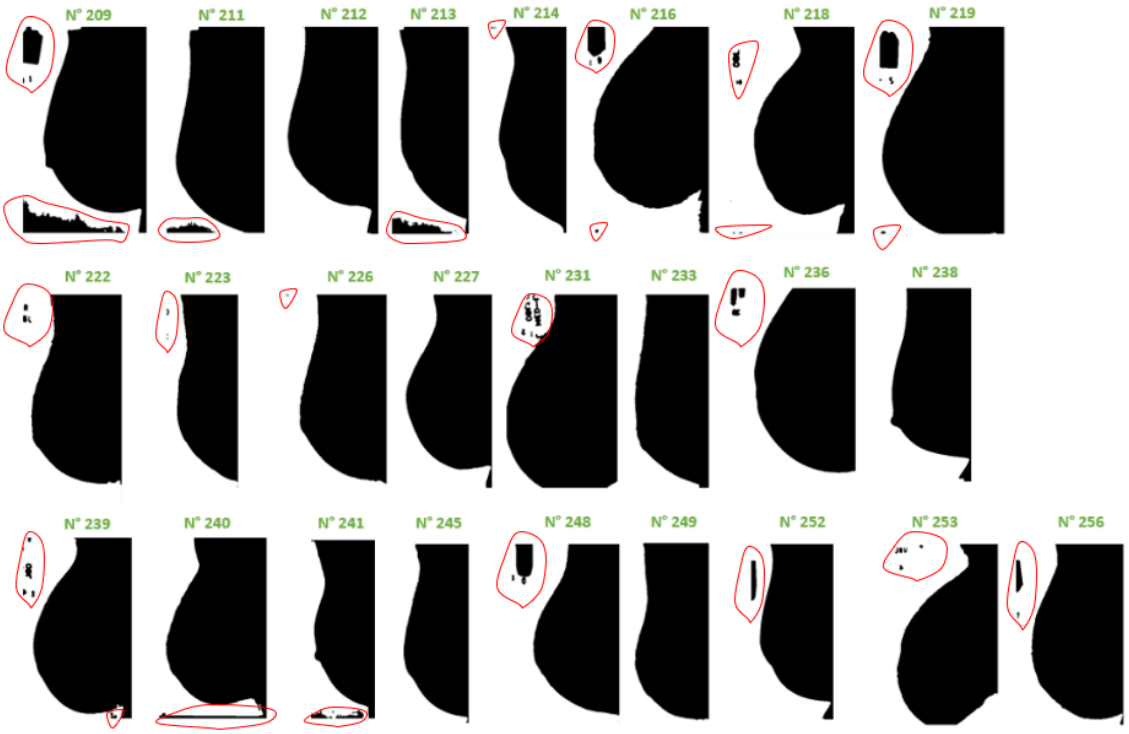


Figura 4.15: Imágenes binarias resultado de aplicar la binarización usando el umbral obtenido mediante la técnica del histograma del gradiente junto con demarcaciones en rojo para los artefactos presentes. Fuente propia

4.2.3.2 Supresión de artefactos

La presencia de artefactos en algunas de las imágenes binarias no permite la correcta separación del fondo y el seno con su tejido pectoral, por ende deben de ser eliminados para poder obtener una máscara que represente única y exclusivamente el seno. Es importante resaltar que gracias a la binarización realizada en la sección anterior, se observa que el seno ocupa el área más grande de la imagen binaria.

Como pudo apreciarse anteriormente, el área que representa el seno junto con el tejido pectoral es mayor a la de los artefactos. Siguiendo esta idea, se realizó la supresión de artefactos en las imágenes binarias utilizando la técnica de Etiquetado de componentes conectados (CCL) con el objetivo de inspeccionar sobre la imagen los componentes conectados, contarlos y finalmente, bajo una regla de decisión simple basada en el área de cada componente, mantener el elemento de mayor área que

corresponde directamente al seno.

La figura 4.16 esquematiza el algoritmo diseñado para remover los artefactos presentes en la imagen de prueba N° 216. Es importante resaltar que la técnica CCL opera sobre la imagen binaria teniendo en cuenta que los elementos a analizar deben de encontrarse representados como píxel blanco (valor 1), por tanto, fue necesario realizar el complemento a la imagen de entrada proveniente de la sección anterior 4.2.3.1. Como se puede visualizar en la figura 4.16, para éste caso específico de la imagen N° 216 la técnica CCL utilizada detectó una cantidad de regiones $L = 6$ $(0, 1, 2, 3, 4$ y $5)$, seguido a esto, se calcula la cantidad de píxeles pertenecientes a cada una de las regiones y la resolución espacial 50 μm por píxel. Posteriormente, conociendo la región con mayor área (perteneciente al seno), se eliminan las otras regiones logrando suprimir los artefactos y elementos presentes, generando así la imagen máscara binaria final. La región 4 contiene 7867303 píxeles siendo la mayor entre las otras regiones. La tabla 4.3 muestra la cantidad de píxeles por cada región L encontradas.

Región *(L)*	0	1	2	3	4	5
Cantidad de píxeles	27.8 %	2 %	0.027 %	0.09 %	70 %	0.037 %

Tabla 4.3: Cantidad de píxeles por región L. Imagen de prueba N° 216.

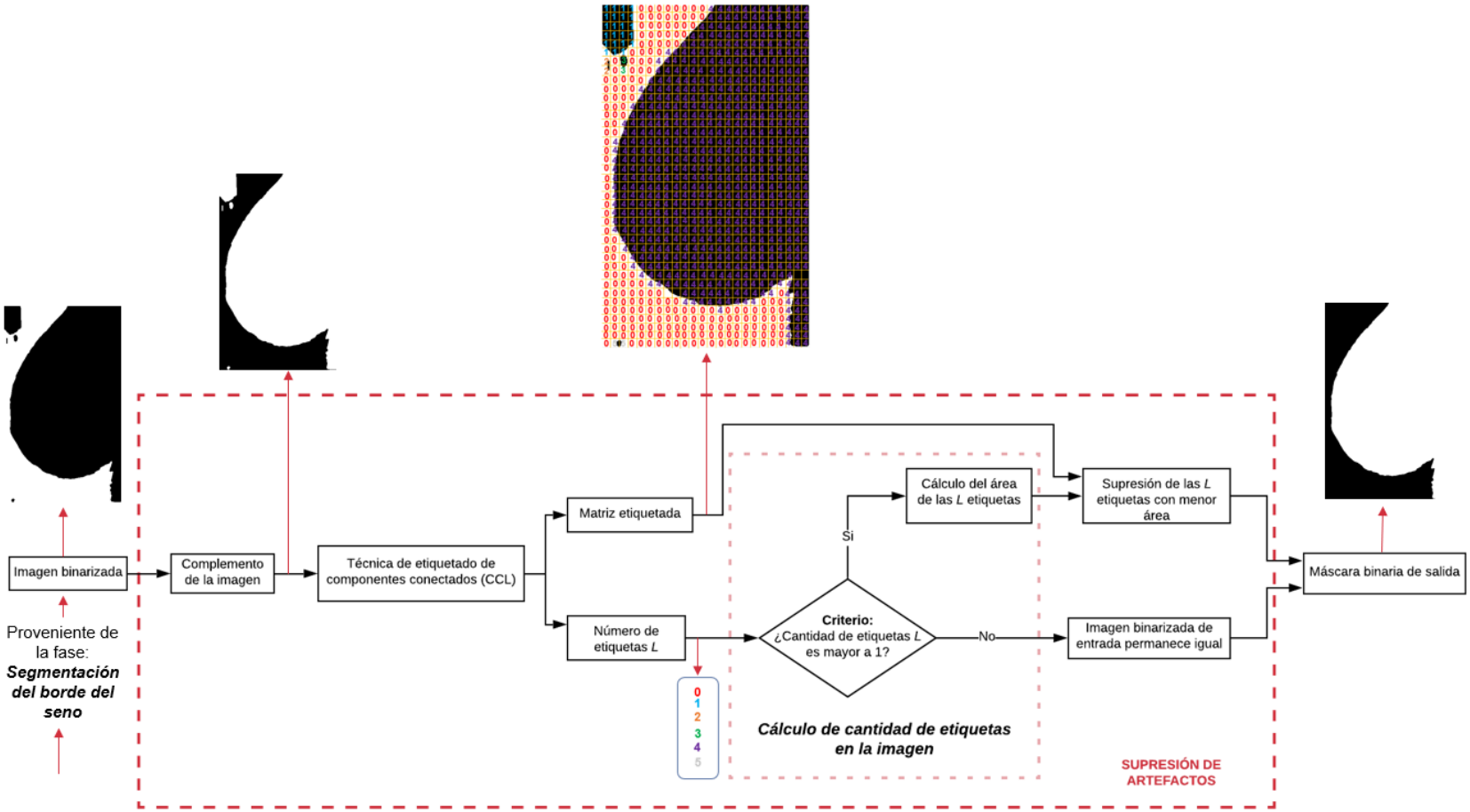


Figura 4.16: Diagrama de bloques del algoritmo para la supresión de artefactos. Imagen de prueba N° 216.

4.2.4 Eliminación del tejido pectoral

Debido a su naturaleza, la región perteneciente al tejido pectoral no tiene relevancia para la inspección de MCs ya que estas no se presentan en esta zona muscular. Sin embargo, la presencia de este en la imagen demanda un costo computacional adicional en el momento de la inspección, por esta razón es necesario segmentarlo para poder evaluar exclusivamente la región del seno, reducir la presencia de falsos positivos en la detección de MCs en esta región y obtener mayor eficiencia en el tiempo de inspección desde el punto de vista computacional.

En la figura 4.17, se muestra la imagen N° 245 como referencia mostrando la región perteneciente al tejido pectoral.

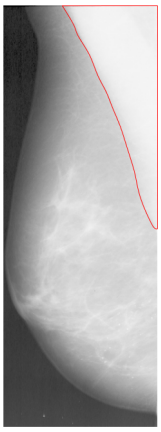


Figura 4.17: Región objetivo a segmentar perteneciente al tejido pectoral. Imagen de prueba N° 245.

La estrategia desarrollada para realización de la segmentación del tejido pectoral se basó en el uso de la BPCS, la figura 4.18 muestra la descomposición en plano de bits para la imagen de prueba N°245 proveniente de la sección 4.2.2.

Dado que el tejido pectoral es una región homogénea, el valor de la intensidad de los píxeles es similar y pertenecen a un rango acotado, se puede apreciar como en el nivel de Plano de bit 5, dicha región se demarca de manera prominente separando el tejido pectoral.

Habiendo analizado el resultado anterior, se procedió a realizar la descomposición

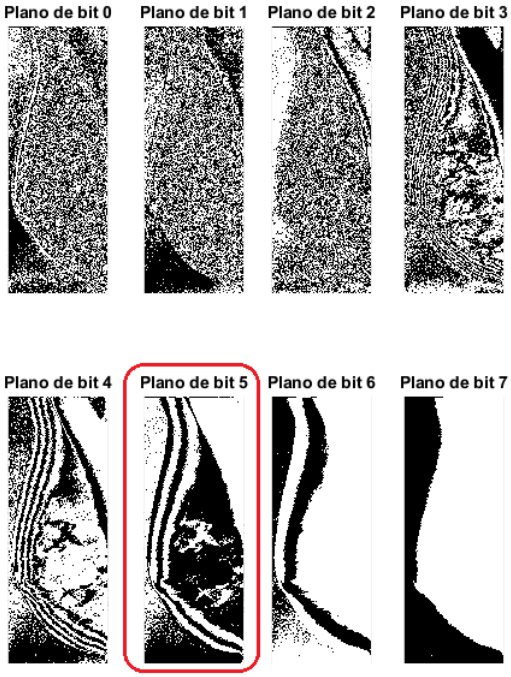


Figura 4.18: Descomposición en planos de bits para imagen de prueba N° 245.

en planos de bits para las 25 imágenes, con el objetivo de visualizar si existe una repetibilidad en la demarcación del tejido pectoral para el Plano de bit 5 de cada una de las imágenes. La inspección visual realizada arrojó que el 80 % de las imágenes presentan una demarcación destacada en la región del tejido pectoral mientras que el 20 % restante no lo presenta debido a una posición de captura no ideal para la identificación del tejido pectoral. Debido a esto, se decidió trabajar con la imagen del plano de bit 5, ya que esta fue la que mejor resaltó el borde del tejido pectoral en aras de diseñar una solución para demarcar y posteriormente suprimir el tejido pectoral.

La figura 4.19 esquematiza la solución desarrollada partiendo del análisis de la imagen del Plano de bit 5 mencionado anteriormente. Inicialmente se calcula el Plano de bit 5 para cada imagen proveniente del resultado obtenido en la sección **Mejora de Contraste** (4.2.2), se realiza una rotación de 270° para poder inspeccionar el borde del tejido pectoral, y posteriormente con los datos obtenidos se realiza un proceso de interpolación para ajustar la curva de una manera suavizada y así poder generar una tendencia del borde del tejido pectoral. Es importante mencionar que el ajuste de la interpolación se realizó probando polinomios de diferente orden y finalmente se estableció el polinomio de orden 3, de la forma mostrada en 4.7, donde

A, B, C y D son la constantes encontradas durante el proceso de interpolación y X los puntos detectados del borde del tejido pectoral.

$$BordeTP = A \cdot X^3 + B \cdot X^2 + C \cdot X + D \tag{4.7}$$

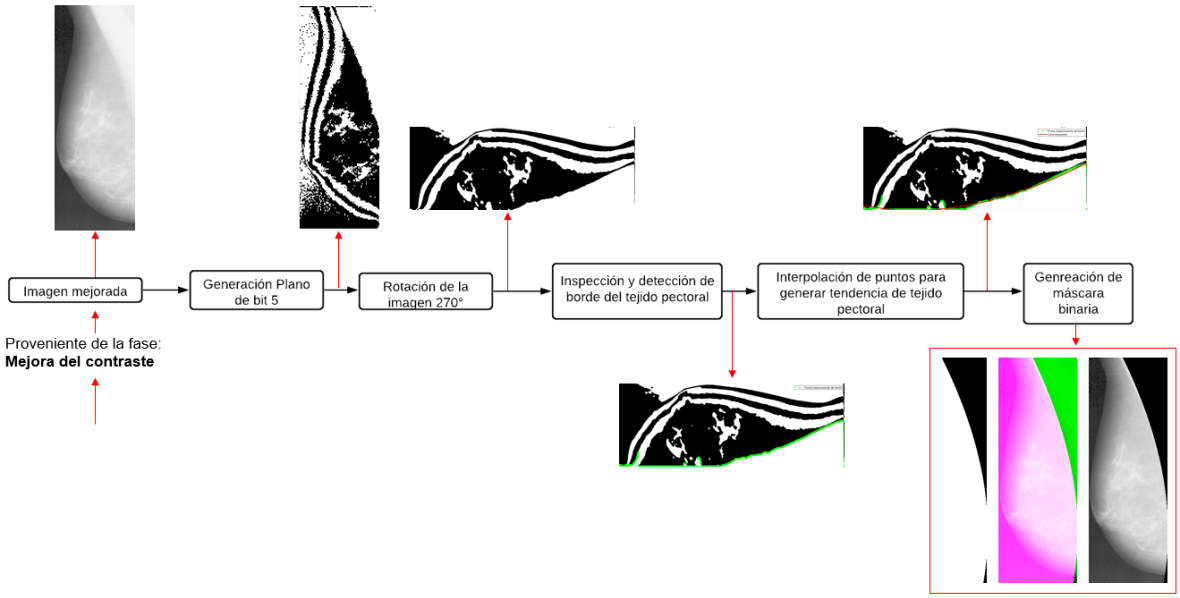


Figura 4.19: Diagrama de bloques del algoritmo para la supresión del tejido pectoral. Imagen de prueba N° 245.

La figura 4.20 presenta los datos pertenecientes a X y el Borde del tejido pectoral interpolado ($BordeTP$) de la ecuación 4.7.

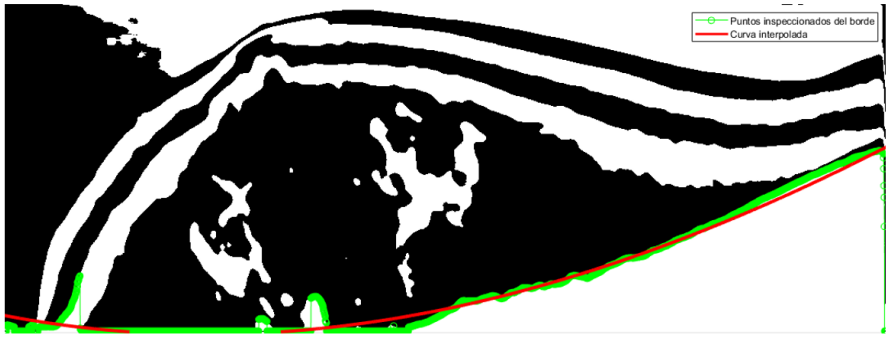


Figura 4.20: Interpolación del borde del tejido pectoral. Imagen de prueba N° 245.

Una vez aplicado el algoritmo anteriormente mencionado, las imágenes han sido procesadas totalmente y se ha logrado llegar a cumplir el objetivo principal:

"Segmentar el seno y tejido pectoral"para posteriormente trabajar únicamente con la región de interés deseada. Cabe resaltar que en algunas imágenes no se obtuvieron los resultados más óptimos debido a múltiples elementos como: exceso de ruido y presencia de etiquetas. Sin embargo, los procedimientos realizados en las secciones anteriores lograron contrarrestar el efecto de dichos elementos y lograr obtener el resultado de segmentación deseado que se presenta para las 25 imágenes en la figura 4.21.

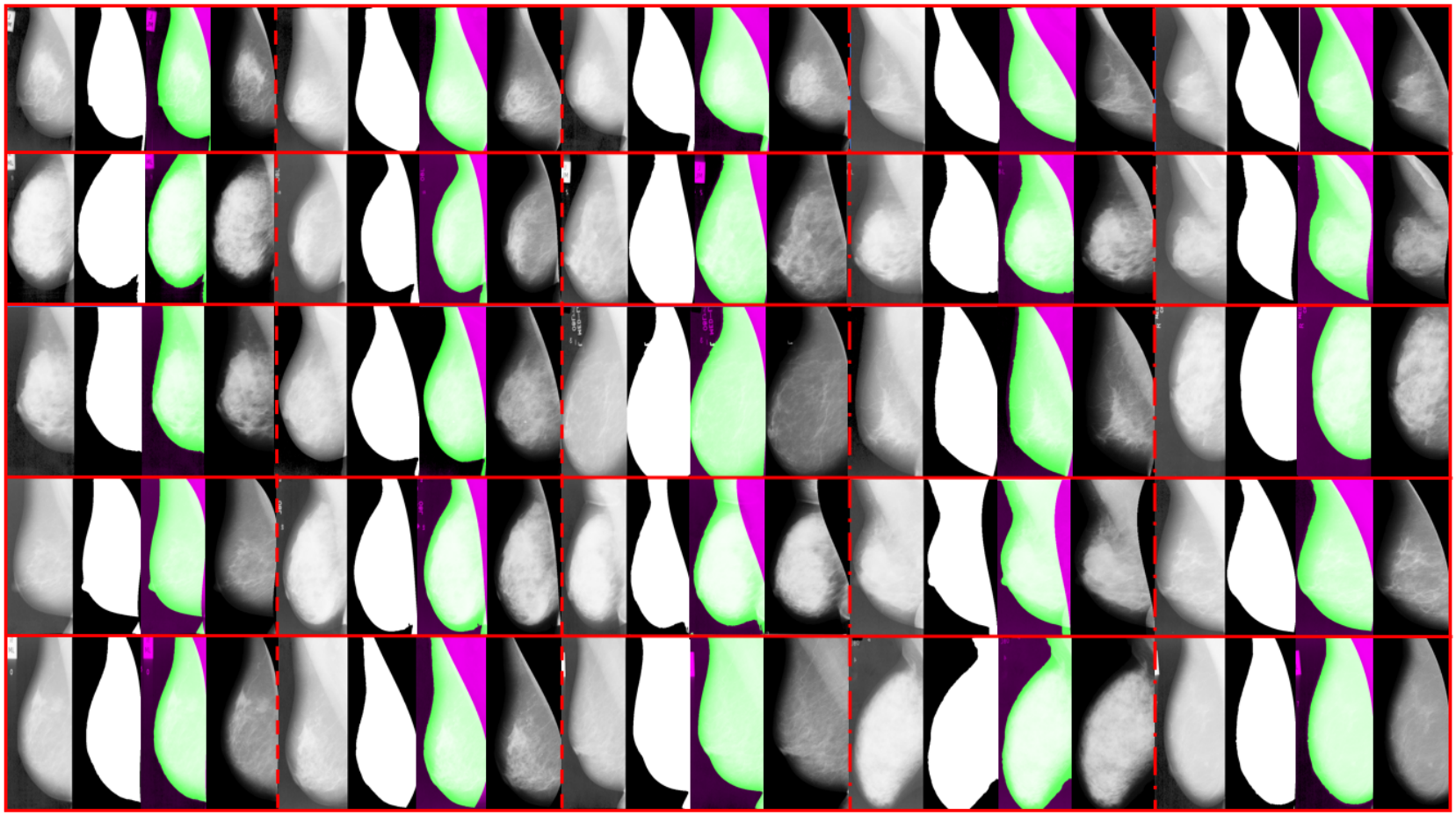
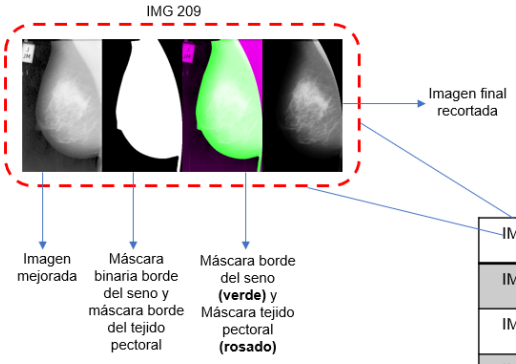


Figura 4.21: Segmentación final borde del seno y tejido pectoral para las 25 imágenes.

Es importante mencionar que la estructura de la imagen presentada en la figura 4.21, esta organizada en forma de mosaico como una matriz de tamaño 5x5 donde cada posición representa una de las 25 imágenes seleccionadas, y a su vez dentro de cada posición del mosaico existen cuatro imágenes: Imagen Mejorada (sección 4.2.2), Máscara Binaria del borde del seno y tejido pectoral (sección 4.2.3 y 4.2.4), Imagen Fusión máscara borde del seno y tejido pectoral y finalmente la imagen original recortada habiendo sido aplicada la máscara binaria del borde del seno y tejido pectoral. La figura 4.22 presenta los índices de las 25 imágenes y su interpretación.



IMG 209	IMG 211	IMG 212	IMG 213	IMG 214
IMG 216	IMG 218	IMG 219	IMG 222	IMG 223
IMG 226	IMG 227	IMG 231	IMG 233	IMG 236
IMG 238	IMG 239	IMG 240	IMG 241	IMG 245
IMG 248	IMG 249	IMG 252	IMG 253	IMG 256

Figura 4.22: Índices imágenes segmentación final borde del seno y tejido pectoral para las 25 imágenes. Ejemplo imagen N°209.

4.3 Constitución de la base de datos para el entrenamiento de las máquinas de aprendizaje

Una vez seleccionadas las 25 mamografías digitales y teniendo en cuenta que la metodología de inspección que se utiliza es mediante la técnica de ventanas deslizantes que se presentará más adelante, se debió seleccionar un tamaño de ventana de inspección cuadrada para la creación de las dos clases a utilizar: (1). región con presencia de MCs y (2). región sana.

Para establecer el tamaño de la ventana de inspección, se realizó una revisión manual de las 25 mamografías para las cuales se midió el diámetro de las microcalcificaciones encontrando así que la microcalcificación más grande presentaba un diámetro de 20 píxeles. Por lo anterior, se estableció una ventana de inspección cuadrada de tamaño

55x55 píxeles, es decir aproximadamente un 36 % mayor que la microcalcificación más grande con el objetivo de asegurar que la ventana de inspección contenga la más grande.

La extracción de las ventanas de inspección se realizó manualmente utilizando las etiquetas previas de MCs, y recortando un área con el tamaño de ventana seleccionada al rededor de las Mcs (ya que el radio provisto en las etiquetas del dataset, corresponde a un área con presencia de MCs, sin embargo también existe una amplia región correspondiente a tegido sano), obteniendo 60 imágenes por clase de interés: (1). región con presencia de MCs y (2). región sana.

4.3.1 Aumento de datos

Debido a que esta cantidad de datos es relativamente pequeña, se realizó un proceso de aumento de datos mediante cuatro transformaciones a las imágenes pertenecientes a cada clase, obteniendo así 240 imágenes nuevas para cada clase en mención, incrementando a 300 imágenes para la clase (1). región con presencia de MCs y 300 imágenes para la clase (2). región sana. Las cuatro transformaciones realizadas se mencionan abajo y por otro lado En la 4.23 se visualiza el proceso en mención.

- **(1). Transformación cuadrática de contraste con una rotación de 90°**: $p_1 = \frac{I(x,y)^2}{255}$; $Ir_1 = rot(p_1, 90°)$.
- **(2). Transformación cúbica de contraste con una rotación de 180°**: $p_2 = \frac{I(x,y)^3}{65535}$; $Ir_2 = rot(p_2, 180°)$.
- **(3). Transformación raíz cuadrada de contraste con una rotación de 270°**: $p_3 = \sqrt{I(x,y) \cdot 255}$; $Ir_3 = rot(p_3, 270°)$.
- **(4). Desenfoque sin rotación**: función *blur* de Matlab con longitud de kernel $w = 1$.

El uso de las transformaciones escogidas (1), (2) y (3) para el aumento de datos se justifica debido a que mediante el análisis visual realizado en las mamografías se pudo

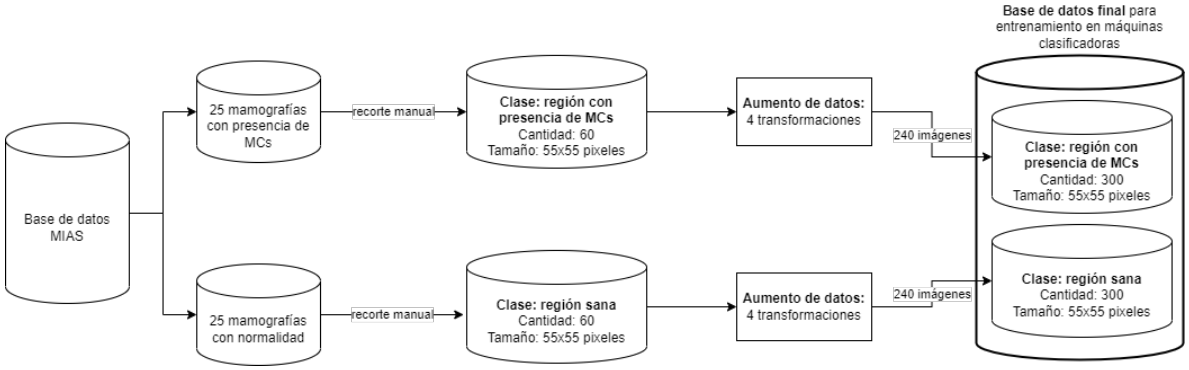


Figura 4.23: Diagrama de bloques del proceso de aumento de datos realizado. Fuente propia

observar que la presencia de MCs puede suceder sin importar el tipo de densidad mamaria la cual visualmente afecta el contraste de la región visualizada en donde se puede apreciar que entre más densidad mamaria el contraste entre la microcalcificación y el fondo es menor donde también se aprecia que se oscurece el fondo, lo que genera una diferenciación más compleja de la microcalcificación, mientras que cuando la densidad mamaría es baja, el contraste entre la microcalcificación y el fondo es mayor apreciando que se aclara el fondo, lo genera una diferenciación más sencilla de la microcalcificación. Así las cosas, las transformaciones (1), (2) y (3) permiten generar una mayor variabilidad de regiones simulando distintos tipos de densidades mamarias.

Las rotaciones realizadas sobre las transformaciones obedecen al objetivo de enriquecer la orientación de la posible microcalcificación y que en el posterior entrenamiento de los clasificadores, éstos tengan la posibilidad de obtener características en diversas posiciones espaciales en la ventana de inspección.

Finalmente, en las mamografías digitales inspeccionadas también se aprecia que algunas microcalcificaciones están bien definidas, otras no, con un efecto de desenfoque. Por lo anterior, se realizó un desenfoque leve mediante la transformación presentada en (4) sobre las ventanas extraídas con el objetivo de enriquecer la base de datos en esta característica en particular.

4.4 Extracción de características

Ahora que se consolidó una base de datos conformada por sub-imágenes extraídas como se mencionó en la etapa de la sección 4.3, se abordarán las técnicas que se utilizaron para extraer la mayor cantidad de características posibles dentro de las ventanas extraídas y que permitan diferenciar las dos clases presentes en el problema (presencia de MCs y sana), esto con el objetivo de consolidar un único vector de características para la posterior clasificación.

Como se mencionó en el marco teórico, normalmente se buscan características que sean similares para objetos que pertenezcan a la misma categoría y diferentes entre categorías. A continuación se describen los métodos utilizados para la extracción de características en la tabla 4.4.

Característica	**Técnica**	**Tipo**	**Alteración**	**Posición**
4 primeros momentos	Manual	Estadística	Ninguna	1 a 4
Entropía, energía, mediana y suavidad	Manual	Estadística	Ninguna	117 a 120
14 momentos de Haralick	Matriz de coocurrencia	Textura	0°,45°, 90° y 135° distancia 1 px	5 a 60
14 momentos de Haralick	Matriz de coocurrencia	Textura	0°, 45°, 90° y 135° distancia 2 px	61 a 116
LRE, GLN, RLN, RP y LGLRE	GLRLM	Textura	Ninguna	121 a 126
LBPU	LBP	Textura	Promedio y desviación de los 20 LBPU centrales	127 y 128
Fractales	SFTA	Textura frecuencial	Ninguna	129 a 149

Tabla 4.4: Resumen de la extracción de características. Fuente propia

Primeramente, se realizó el cálculo de las características estadísticas, más concretamente los 4 primeros momentos estadísticos, esto debido a que trabajos similares implementan estas características regularmente. Así mismo, se calculan la entropía, energía, mediana y suavidad. Posteriormente se implementó el método de Haralick [Haralick et al., 1973] variando los ángulos y su distancia, con el fin de extraer características de textura variadas, las cuales resultan de gran ayuda en el reconocimiento de objetos. Después, se calculan seis de las características que

proporcionan la matriz GLRLM. Luego de esto, se implementó un algoritmo que calcula la matriz correspondiente al código LBP de las 600 imágenes con el fin de obtener su histograma y así obtener las 59 características de los patrones uniformes utilizando el método de Patrones binarios locales uniformes (LBPU), de estos 59 se extraen los 20 centrales y se calcula su promedio y desviación estándar, obteniendo así, dos características adicionales para el vector. Finalmente, se utilizó el algoritmo de SFTA [1] para calcular las últimas 20 características.

Una vez aplicados los métodos seleccionados para la extracción de características, se obtuvo un vector de dimensión [1 x 149] para cada imagen, siendo 149 el número de características extraídas por las diferentes técnicas y distribuidas como se observa en la figura 4.24.

Vector de Características
[1-4] ⟶ Momentos estadísticos
[5-60] ⟶ Haralick con distancia de 1 px para los cuatro ángulos.
[61-116] ⟶ Haralick con distancia de 2 px para los cuatro ángulos.
[117-120] ⟶ Otras características estadísticas
[121-126] ⟶ GLRLM
[127-128] ⟶ LBPU
[129-149] ⟶ SFTA

Figura 4.24: Descripción del vector de características obtenido por cada imagen. Fuente propia

4.4.1 Eliminación de datos atípicos

La matriz de características obtenida contiene ciertos datos de tipo Not a Number (NaN), esto se debe a que para ciertas imágenes, el cálculo de las características de textura se indetermina por cuestiones de homogeneidad en sus niveles de gris. En esta sección se realizó la búsqueda de elementos tipo NaN, con el fin de descartar los vectores que contengan dicho resultado. Al realizar esta búsqueda se encontraron ocho muestras

[1] Alceu Costa (2022). alceufc/sfta (https://github.com/alceufc/sfta), GitHub. Retrieved May 19, 2022.

con presencia de NaN para la clase con presencia de MCs y de 48 para la clase sana. Estos NaN, se presentaron en las características de textura, es normal que en la clase sana exista mayor cantidad debido a que la textura representa cambios de alta frecuencia en la imagen y las secciones de la mama donde no hay presencia de MCs suelen ser mayormente planas.

4.4.2 Normalización de las características

Una vez descartado las muestras con NaN, se obtuvo un dataset desbalanceado con 292 muestras para la clase con presencia de MCs y 252 para la clase sana. A estas muestras se les aplicó cuatro tipos de normalización: de 0 a 1, de -1 a 1, estandarización y una no lineal, softmax con un $r = 0.4$. Lo anterior con el fin de validar el resultado de los algoritmos frente a la variación de los rangos y comportamiento de las características.

4.5 Selección de características

Gracias a la extracción de características, se obtuvo cinco matrices de 149 características, uno por cada normalización y una sin normalizar. Sin embargo, es común que se realice una selección de características con el fin de escoger las más discriminantes y a su vez, reducir la dimensionalidad del problema.

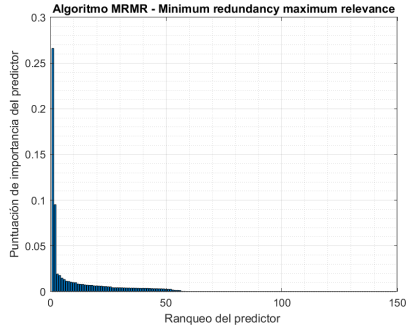
Para la selección de características se implementaron los cuatro métodos mencionados en el capítulo 3. Estos métodos proporcionan puntajes y rankings de las diferentes características teniendo en cuenta distintos valores estadísticos para determinar cuales de las 149 son las más discriminantes y, por ende, más importantes para la etapa de clasificación.

Se aplicaron los cuatro algoritmos de ponderación de características a las cinco matrices, obteniendo una ponderación por cada algoritmo y por cada normalización. Este esquema de pruebas da como resultado 20 vectores de ponderación: cinco para cada algoritmo con diferentes normalizaciones.

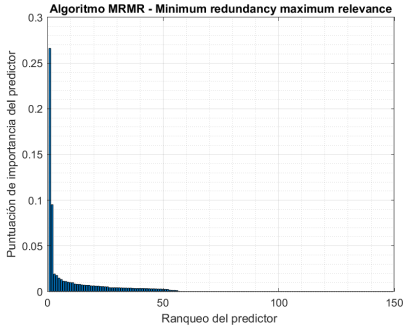
Los resultados de las ponderaciones con el método de MRMR se observan en las

figuras 4.25, 4.26 y 4.27.

Se puede apreciar que las normalizaciones no afectan mucho en la ponderación para este algoritmo. Adicionalmente, se puede apreciar que gran parte de la información que proveen las características se concentra entre las 10-20 primeras en el ranking provisto por MRMR.

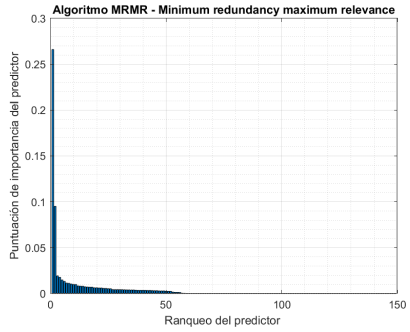


(a) Ponderación del método MRMR con datos sin normalizar.

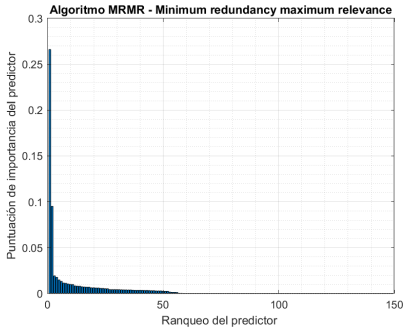


(b) Ponderación del método MRMR con datos normalizados de 0 a 1.

Figura 4.25



(a) Ponderación del método MRMR con datos normalizados de -1 a 1.



(b) Ponderación del método MRMR con datos estandarizados.

Figura 4.26

Ahora, los resultados de las ponderaciones con el algoritmo de fscchi2 se observan en las figuras 4.28, 4.29 y 4.30.

Se puede apreciar que las normalizaciones tienen implicaciones mínimas en la ponderación para este algoritmo. Además, la mayor parte de la información se concentra al rededor de las primeras 30-40 primeras características. Para el algoritmo

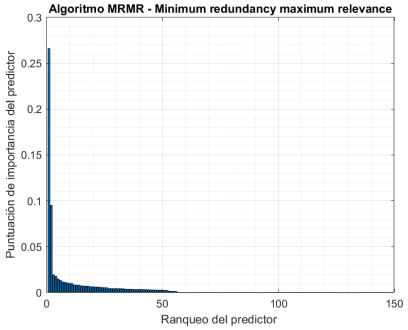
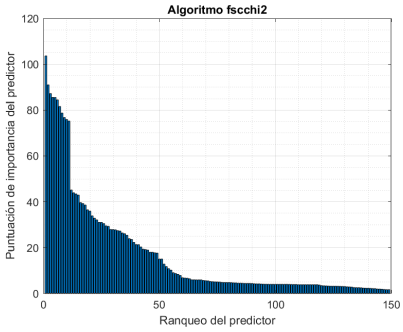
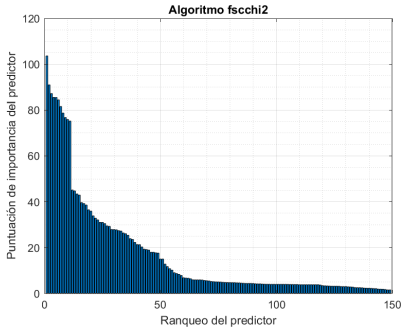


Figura 4.27: Ponderación del método MRMR con datos normalizados con la función soft-max.

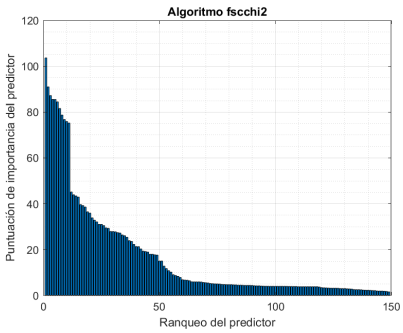


(a) Ponderación del método fscchi2 con datos sin normalizar.

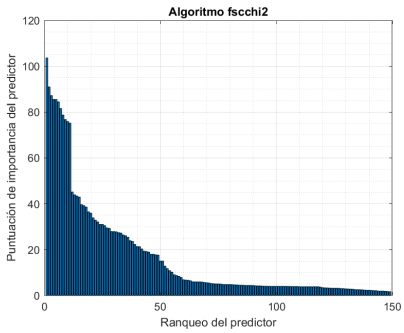


(b) Ponderación del método fscchi2 con datos normalizados de 0 a 1.

Figura 4.28



(a) Ponderación del método fscchi2 con datos normalizados de -1 a 1.



(b) Ponderación del método fscchi2 con datos estandarizados.

Figura 4.29

basado en NCA, se tomaron los pesos resultantes del entrenamiento y se realizó la ponderación manualmente para obtener resultados comparables con otros métodos.

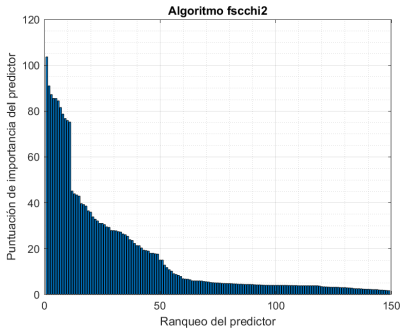
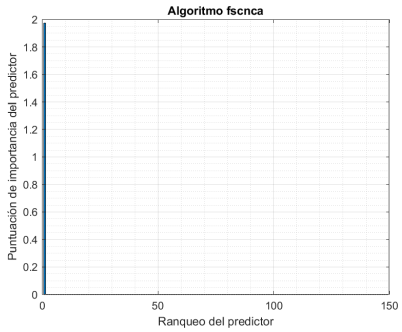
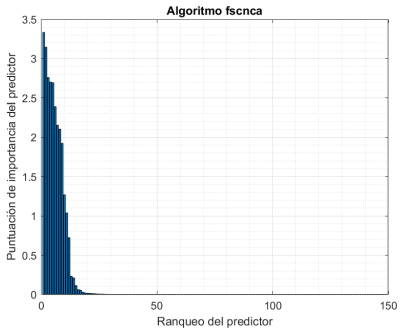


Figura 4.30: Ponderación del método fscchi2 con datos normalizados con la función soft-max.

Los resultados de las ponderaciones con el algoritmo de fscnca se observan en las figuras 4.31, 4.32 y 4.33.

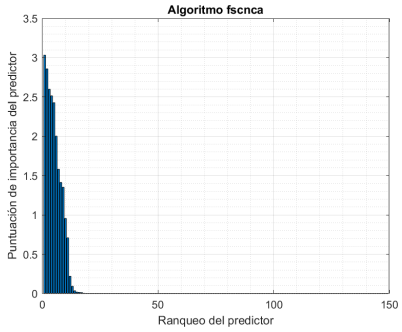


(a) Ponderación del método fscnca con datos sin normalizar.

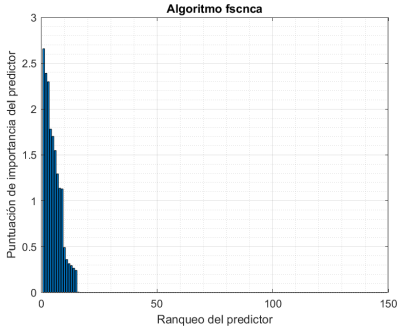


(b) Ponderación del método fscnca con datos normalizados de 0 a 1.

Figura 4.31



(a) Ponderación del método fscnca con datos normalizados de -1 a 1.



(b) Ponderación del método fscnca con datos estandarizados.

Figura 4.32

También cabe resaltar, que los pesos resultantes para este algoritmo serán 0 o muy

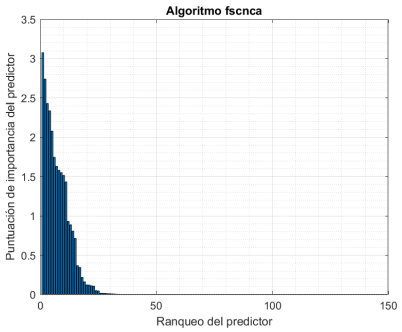
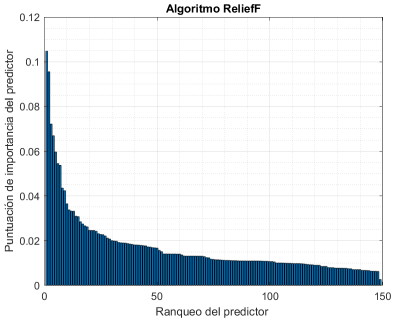


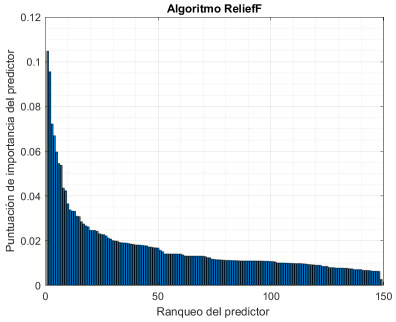
Figura 4.33: Ponderación del método fscnca con datos normalizados con la función soft-max.

cercanos a 0 para las características irrelevantes. Como se observa en la figura 4.31, para fscnca, la normalización de los datos es relevante para que la ponderación sea adecuada. Adicionalmente, se observa que la mayor concentración de información se presenta al rededor del 10 % de las características.

Por último, las ponderaciones resultantes de aplicar el método Relief se aprecian en las figuras 4.34, 4.35 y 4.36.



(a) Ponderación del método relief con datos sin normalizar.

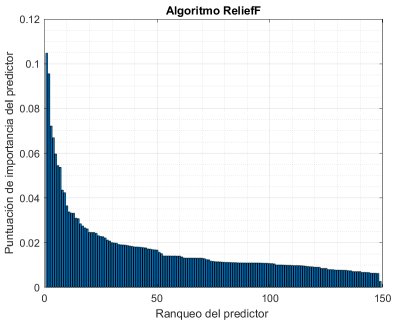


(b) Ponderación del método de relief con datos normalizados de 0 a 1.

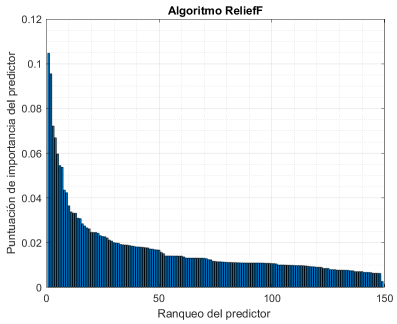
Figura 4.34

Se puede apreciar que la normalización tampoco afecta en gran medida la distribución de la ponderación para este algoritmo.

Ahora, para tener una mejor apreciación de las mejores características para entrenar los modelos, se realizó una tabla de ponderaciones para cada normalización, tomando en cuenta su posición en las ponderaciones de los cuatro algoritmos. Lo



(a) Ponderación del método de relief con datos normalizados de -1 a 1.



(b) Ponderación del método de relief con datos estandarizados.

Figura 4.35

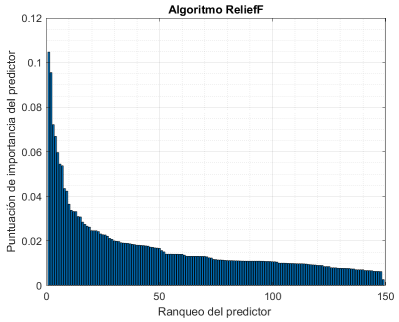


Figura 4.36: Ponderación del método de relief con datos normalizados con la función soft-max.

anterior proporciona como resultado cinco vectores de ponderación (uno para datos sin normalizar y uno por cada normalización realizada), de los cuales se observan las 15 primeras posiciones en la tabla 4.5, cabe resaltar que la tabla completa se encuentra en el apartado de apéndices.

La ponderación para la tabla 4.5 se realizó siguiendo la ecuación 4.8

$$FP = RP_1 + RP_2 + RP_3 + RP_4 \tag{4.8}$$

Donde FP es la ponderación final y RP_i es la posición en el ranking del algoritmo iésimo. Cabe resaltar que esta operación se repite por cada tipo de normalización.

Gracias a las gráficas de los *rankings* proporcionados por los distintos métodos, se llegó a la conclusión que el realizar la normalización de los datos no influye

	N° de característica				
Rank	**Sin normalizar**	**De 0 a 1**	**De -1 a 1**	**Estandarización**	**Soft-max**
1	126	126	126	3	3
2	3	138	138	126	126
3	140	127	127	127	127
4	125	125	3	138	140
5	149	140	140	140	138
6	138	3	125	124	121
7	4	149	124	125	125
8	124	121	149	121	124
9	143	124	121	149	122
10	147	4	4	4	4
11	139	30	102	139	149
12	137	102	139	137	46
13	127	139	147	128	102
14	121	32	137	147	147
15	2	137	30	30	60

Tabla 4.5: Tabla de ponderaciones de las 15 mejores características según el tipo de normalización utilizando los cuatro algoritmos.

significativamente en las técnicas de selección de características, sin embargo, dado que los índices de las mejores características variaron en algunas normalizaciones, se decidió extraer un vector nuevo procedente de la combinación de los *rankings* de la tabla 4.5.

Por último se realizó una vez más una tabla de ponderaciones utilizando los cinco vectores obtenidos previamente de la tabla 4.5, obteniendo un único vector de ponderaciones, el vector se organizó en la tabla 4.6.

La ponderación para la tabla 4.6 se realizó siguiendo la ecuación 4.9

$$FP = RP_1 + RP_2 + RP_3 + RP_4 + RP_5 \tag{4.9}$$

Donde FP es la ponderación final y RP_i es la posición en el ranking de la tabla 4.5 para la iésima normalización.

Así mismo, en la tabla 4.6 se mencionan las 15 mejores características, esto se

Rank	N° de característica
1	126
2	3
3	138
4	140
5	127
6	125
7	124
8	149
9	121
10	4
11	139
12	137
13	147
14	102
15	30

Tabla 4.6: Tabla de ponderaciones de las 15 mejores características según los cinco tipos de normalización y los cuatro algoritmos implementados.

fundamenta en el hecho que la mayor parte de la información, de acuerdo a las técnicas de ranking implementadas, se encontraba cerca del 10 % del total de las características iniciales (149) y esto se pudo apreciar en las distintas gráficas de las ponderaciones por cada método. Sin embargo, la tabla 4.6 se extiende en la sección de apéndices.

4.6 Entrenamiento

4.6.1 Protocolo de entrenamiento

Con el fin de realizar distintas pruebas, se realizó la creación de 100 universos de entrenamiento y testeo, distribuidos aleatoriamente en un 70 % para entrenamiento y 30 % para testeo y posteriormente, se implementó un protocolo de entrenamiento con cuatro tipos de clasificadores: RNA, SVM, K-NN y NB, de estos entrenamientos, se seleccionó las 10 redes de cada clasificador con mejor desempeño en la métrica de exactitud para posteriormente someterlos a un análisis ANOVA.

El esquema de pruebas se realizó para los dos vectores de características (columna

"sin normalizar"de la tabla 4.5 y la columna de la tabla 4.6), esto con el fin de validar la segunda ponderación implementada en la sección anterior. A continuación se explica el protocolo de pruebas para cada modelo:

4.6.1.1 ANN

Para la red neuronal se tomó cada uno de los universos y se entrenó 20 modelos para cada uno, variando el número de neuronas en la capa oculta (entre 10 y 30 neuronas). Con un total de 2000 modelos entrenados, se seleccionó el que tuvo el mejor desempeño según la métrica de exactitud y teniendo en cuenta el valor p de significancia provisto por el análisis ANOVA.

4.6.1.2 SVM

Para la SVM se crearon tres modelos para cada universo, variando el tipo de kernel de la máquina (Lineal, Gaussiano y Polinomial). Con un total de 300 modelos entrenados, se seleccionó el que tuvo el mejor desempeño según la métrica de exactitud y teniendo en cuenta el valor p de significancia provisto por el análisis ANOVA.

4.6.1.3 KNN

Para el K-NN se crearon 18 modelos para cada universo, variando el número de vecinos del modelo (k desde 2 a 20). Con un total de 1800 modelos entrenados, se seleccionó el que tuvo el mejor desempeño según la métrica de exactitud y teniendo en cuenta el valor p de significancia provisto por el análisis ANOVA.

4.6.1.4 BY

Para el NB, primero se realizó una prueba de gaussianidad a las 15 características, incluyendo tres test distintos (Lilliefors, Anderson-Darling y Kolmogorov-Smirnov). Posteriormente, se entrenó un modelo para cada universo. Con un total de 100 modelos entrenados, se seleccionó el que tuvo el mejor desempeño según la métrica de

exactitud y teniendo en cuenta el valor p de significancia provisto por el análisis ANOVA.

4.7 Inferencia y post-validación

Para la inferencia, se utilizó una metodología basada en un deslizamiento de ventana, tomando el mismo tamaño de la ventana que se implementó para la extracción de la región de interés en el dataset de entrenamiento.

El algoritmo realiza un barrido horizontal y vertical de la ventana sobre la imagen que a la que se quiere realizar la inferencia, aplicando un porcentaje de traslape del 18 % aproximadamente. Luego, se calcula el vector de características para cada una de las regiones y se almacena para obtener una predicción utilizando los modelos entrenados en la fase anterior.

4.7.1 Post validación

Una vez se tienen las predicciones para cada ventana, se realiza una etapa de post validación para que los resultados de la inferencia sean más confiables.

4.7.1.1 Post - análisis 1

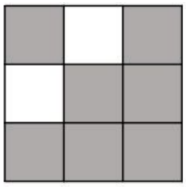
Para este post análisis, primero se agrupan en matrices de 3 x 3, las regiones ya con su inferencia y luego se aplica la condición:

Si al menos siete de las nueve regiones pertenecen a la misma clase, toda la región post análisis (Matriz 3x3) es clasificada como perteneciente a dicha clase, tal y como se observa en la figura 4.37a.

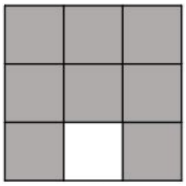
4.7.1.2 Post - análisis 2

Para este post análisis, primero se agrupan en matrices de 3 x 3, las regiones ya con su inferencia y luego se aplica la condición:

Si al menos ocho de las nueve regiones pertenecen a la misma clase, toda la región post análisis (Matriz 3x3) es clasificada como perteneciente a dicha clase, tal y como se observa en la figura 4.37b.



(a) Ejemplo de Post-análisis 1.



(b) Ejemplo de Post-análisis 2.

Figura 4.37: Ejemplos de los dos post-análisis desarrollados. [Fuente: Elaboración propia].

5. Pruebas y resultados

Para este capítulo se realizó un protocolo de pruebas que midan y evalúen el desempeño de la metodología empleada para la detección automática de microcalcificaciones en mamografías digitales. Antes de implementar el protocolo, se seleccionan tres mamografías digitales donde previamente se conoce la presencia de microcalcificaciones, a estas imágenes se les aplica el preprocesamiento mencionado en la sección 4.2 y se eliminan los artefactos y el pectoral presente en la mamografía. Gracias a esto, la zona a inspeccionar en la etapa de inferencia se limita a la mama. Una vez realizado el preprocesamiento a cada imagen se procede a implementar el protocolo de pruebas descrito a continuación.

5.1 Protocolo de pruebas

5.1.1 Método de enventanado

Debido a que las técnicas de extracción de características fueron implementadas en regiones de las imágenes denominadas ventanas, con un tamaño de $55x55$ píxeles, para las pruebas se deben extraer las características locales en regiones de la misma dimensión. Para generar estas regiones, se realizó un enventanado de la imagen con dimensiones de $55x55$ y un porcentaje de traslape del 82 %. Gracias a esto, se obtienen más de 100.000 imágenes a las cuales se les puede extraer las características descritas en la figura 4.24.

5.1.2 Validación de la segmentación del seno y el tejido pectoral

Dado que la base de datos MIAS no presenta información para realizar una comparación sobre la segmentación del borde del seno y el tejido pectoral, se utilizó una herramienta de etiquetado manual para realizar las imágenes objetivo y así poder determinar, mediante métricas de similaridad entre imágenes, el desempeño de la solución planteada.

La figura 5.1 presenta un ejemplo en del etiquetado y la máscara final obtenida.

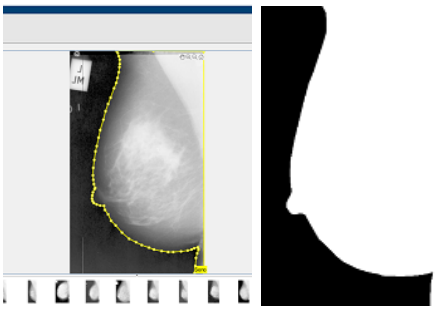


Figura 5.1: Proceso de etiquetado manual del borde del seno (izquierda) y la máscara obtenida manualmente (derecha) para la imagen N° 209 base de datos MIAS.

Con el objetivo de medir el desempeño de la técnica utilizada para la obtención de la máscara binaria, se utilizaron dos métricas para evaluar el proceso de segmentación del seno.

La métricas coeficiente de Sorensen-Dice y el índice de Jaccard fueron calculadas para las 25 imágenes usando las máscaras binarias realizadas anteriormente como imágenes de referencia. y las imágenes obtenidas con la técnica CCL ilustrada en la figura 4.16. La figura 5.2 presenta el diagrama de cajas comparando los resultado para las dos técnicas.

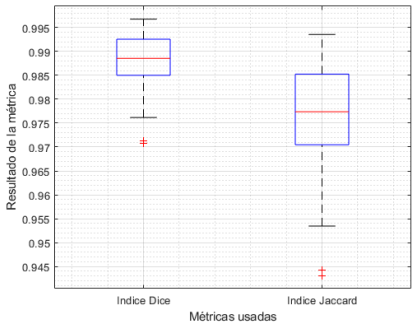


Figura 5.2: Diagrama de cajas resultados coeficiente Sorensen-Dice e Índice de Jaccard para segmentación del seno.

De la figura 5.2, se puede percatar como el coeficiente de Sorensen-Dice y el índice de Jaccard reportan resultados con promedio 0.9874 y 0.9752 respectivamente, lo que indica que el proceso de segmentación para las 25 imágenes ha sido exitoso. Algunos casos atípicos registran valores cercanos al 0.94, sin embargo sigue siendo alto la similaridad

encontrada entre la imagen de referencia y la obtenida mediante la técnica presentada anteriormente.

5.1.3 Descarte de la zona del pectoral y externa a la mama

Una vez extraídas las ventanas, se procede a realizar un procesamiento previo a la extracción de características. Gracias a la eliminación de los artefactos y el pectoral, se obtiene un fondo completamente negro y la mama en escala de grises, aprovechando esta particularidad, se realiza un algoritmo para calcular el promedio de intensidad de grises para cada ventana, así se determinan fácilmente las regiones que pertenecen al fondo negro para descartarlas. Por otro lado, también se encuentran las ventanas que corresponden a la zona de la mama y por ende, es la región de interés para la extracción de características.

5.1.4 Extracción de características

Posteriormente, se realiza la extracción de características a las ventanas que pasaron por el algoritmo mencionado anteriormente, se calculan las 149 características mencionadas en la figura 4.24 y después se seleccionan las 15 mejores según la tabla 4.6. Este vector es almacenado en una matriz cuya posición está relacionada con la ventana y su ubicación espacial en la imagen original.

5.1.5 Entrenamiento utilizando el vector de características de la tabla 4.5 (vector de características 1)

5.1.5.1 Prueba de Gaussianidad

Para conocer si las características tienen un comportamiento Gaussiano se evaluó el resultado de tres test de gaussianidad, el de Lilliefors, Anderson-Darling y por último, el de Kolmogorov-Smirnov, una vez evaluadas las características se excluyen del vector final las que fueron rechazadas por los 3 test para alguna de las dos clases. Los resultados de los test para el vector de características 1 se observan en la tabla 5.1.

Feature	Lilliefors		Anderson		Kolmogorov	
1	1	0	1	0	0	0
2	1	1	1	1	1	0
3	1	1	1	1	1	0
4	1	1	1	1	0	0
5	1	0	1	0	1	0
6	0	1	1	0	0	0
7	1	0	1	0	0	0
8	1	1	1	1	1	1
9	1	0	1	0	0	0
10	1	1	1	1	1	0
11	1	1	1	1	1	0
12	1	1	1	1	0	0
13	1	1	1	1	1	1
14	1	1	1	1	1	1
15	1	1	1	1	1	1

Tabla 5.1: Resultados de los test de gaussianidad de Lilliefors, Anderson-Darling y Kolmogorov-Smirnov

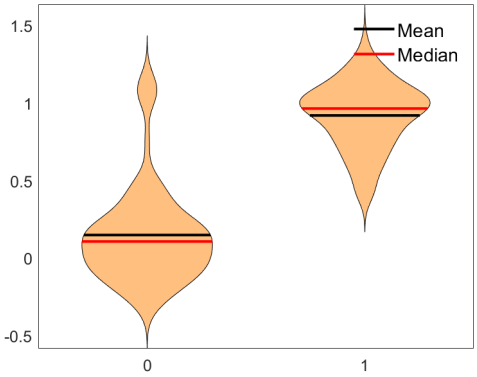
5.1.5.2 ANOVA

Para el análisis de varianza, se tomó un universo de datos diferente al de los entrenamientos, se realizó la predicción con cada clasificador y se comparó con sus respectivas etiquetas. Posteriormente se realizó el diagrama de violín para cada modelo, teniendo así una idea de la distribución de las predicciones.

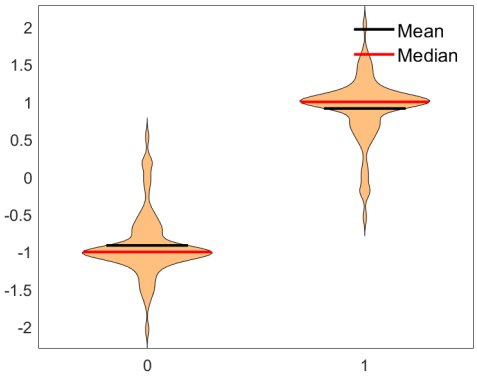
Por ultimo, los mejores modelos (con un mejor nivel de significancia p) por cada clasificador fueron condensados en la tabla 5.2, mostrando los parámetros utilizados según el protocolo descrito anteriormente.

Model	Neurons	k	kernel	Features	Universe	Acc
ANN	24	N/A	N/A	15	5	97.53 %
SVM	N/A	N/A	Gaussian	15	6	95.68 %
KNN	N/A	7	N/A	15	91	95.68 %
BY	N/A	N/A	N/A	6	53	80.25 %

Tabla 5.2: Resultados del protocolo de entrenamiento.

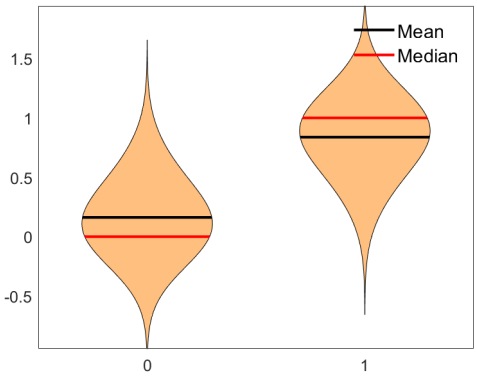


(a) Diagrama de violín para el mejor modelo de ANN.

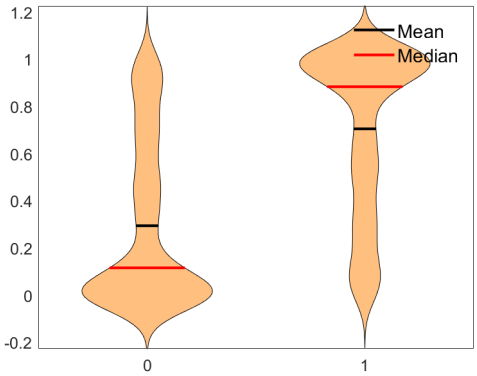


(b) Diagrama de violín para el mejor modelo de SVM.

Figura 5.3: Diagramas de violín para el primer y segundo modelo. [Fuente: Elaboración propia].



(a) Diagrama de violín para el mejor modelo de KNN.



(b) Diagrama de violín para el mejor modelo de Bayes.

Figura 5.4: Diagramas de violín para el tercer y cuarto modelo. [Fuente: Elaboración propia].

5.1.6 Entrenamiento utilizando el vector de características de la tabla 4.6 (vector de características 2)

5.1.6.1 Prueba de Gaussianidad

Para conocer si las características tienen un comportamiento Gaussiano se evaluó el resultado de tres test de gaussianidad, el de Lilliefors, Anderson-Darling y por último, el de Kolmogorov-Smirnov, una vez evaluadas las características se excluyen del vector final las que fueron rechazadas por los 3 test para alguna de las dos clases. Los resultados

de los test para el vector de características 2 se observan en la tabla 5.3.

Feature	**Lilliefors**		**Anderson**		**Kolmogorov**	
1	1	0	1	0	0	0
2	1	1	1	1	1	0
3	1	0	1	0	1	0
4	1	1	1	1	1	0
5	1	1	1	1	1	1
6	1	1	1	1	0	0
7	1	1	1	1	1	1
8	0	1	1	0	0	0
9	1	1	1	1	1	1
10	1	0	1	0	0	0
11	1	1	1	1	1	0
12	1	1	1	1	0	0
13	1	1	1	1	1	0
14	1	1	1	1	0	0
15	0	1	1	1	0	1

Tabla 5.3: Resultados de los test de gaussianidad de Lilliefors, Anderson-Darling y Kolmogorov-Smirnov para el vector de características 2

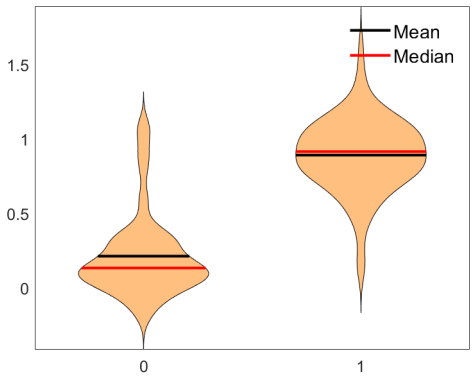
5.1.6.2 ANOVA

Para el análisis de varianza, se tomó un universo de datos diferente al de los entrenamientos, se realizó la predicción con cada clasificador y se comparó con sus respectivas etiquetas. Posteriormente se realizó el diagrama de violín para cada modelo, teniendo así una mejor aproximación sobre la distribución de las predicciones.

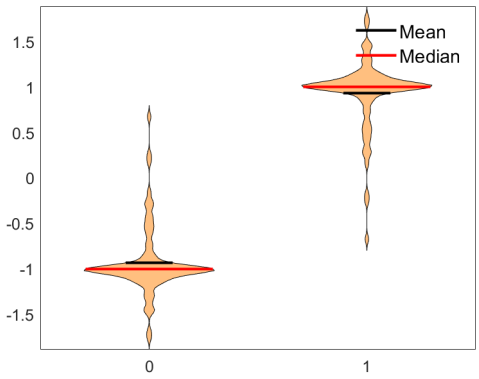
Por ultimo, los mejores modelos (con un mejor nivel de significancia p) por cada clasificador fueron condensados en la tabla 5.4, mostrando los parámetros utilizados según el protocolo descrito anteriormente.

Model	**Neurons**	**k**	**kernel**	**Features**	**Universe**	**Acc**
ANN	23	N/A	N/A	15	31	95.68 %
SVM	N/A	N/A	Gaussian	15	5	95.06 %
KNN	N/A	7	N/A	15	12	95.06 %
BY	N/A	N/A	N/A	6	53	80.86 %

Tabla 5.4: Resultados del protocolo de entrenamiento.

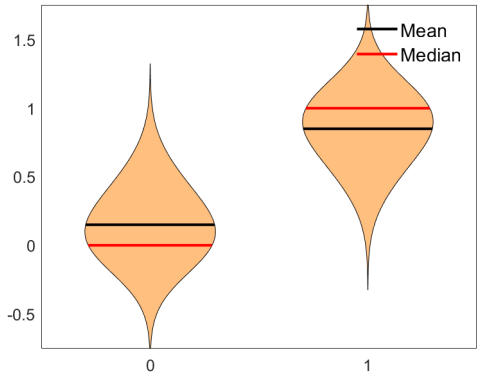


(a) Diagrama de violín para el mejor modelo de ANN.

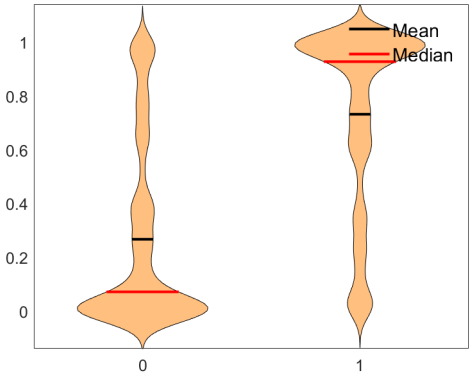


(b) Diagrama de violín para el mejor modelo de SVM.

Figura 5.5: Diagramas de violín para el primer y segundo modelo. [Fuente: Elaboración propia].



(a) Diagrama de violín para el mejor modelo de KNN.



(b) Diagrama de violín para el mejor modelo de Bayes.

Figura 5.6: Diagramas de violín para el tercer y cuarto modelo. [Fuente: Elaboración propia].

5.1.7 Predicción

Teniendo el vector de 15 características para cada ventana, se procede a realizar la predicción con los cuatro modelos mencionados en la tabla 5.2. Gracias a los modelos pre entrenados, simplemente se necesita el vector de características para esta realizar la predicción.

5.1.7.1 Opción de rechazo

Como es de esperarse, la salida de los clasificadores son dos números decimales y_0 y y_1 cuya mayor cercanía a 1 significa una mayor probabilidad de pertenecer a la clase ω_0 y ω_1 respectivamente. Sin embargo, es común establecer una tercera opción (opción de rechazo) para los clasificadores, esto permite a la máquina desistir de la obligación de determinar la clase correspondiente si un dato nuevo está fuera de lo común y no se tiene la información necesaria para determinar la clase. Esto disminuye la tasa de falsos positivos e incrementa la precisión para cada clase ya que las predicciones realizadas tienen una confianza mucho mayor.

Para este problema, se implementó una opción de rechazo con un margen de error del 15 %, permitiendo el paso de las predicciones que se encuentren dentro del umbral de rechazo establecido.

5.1.8 Inferencia sobre la imagen original

Para tener una mejor visualización de los resultados, se decidió crear recuadros de distintos colores para cada clase sobre la imagen original utilizando las coordenadas de las ventanas y su posición conocida en la matriz de predicciones.

5.1.8.1 Post validación nivel 1

Posteriormente, se implementó el análisis mencionado en 4.7.1.1, estableciendo una tasa de 7/9 regiones clasificadas en la misma clase para determinar la clase final post validada de nivel 1.

5.1.8.2 Post validación nivel 2

Así mismo, se implementó el análisis mencionado en 4.7.1.2, estableciendo una tasa de 8/9 regiones clasificadas en la misma clase para determinar la clase final post validada de nivel 2.

5.2 Resultados

5.2.1 Imagen 209

La primera imagen seleccionada para la aplicación del protocolo de pruebas fue la 209 y se observa en la figura 5.7.

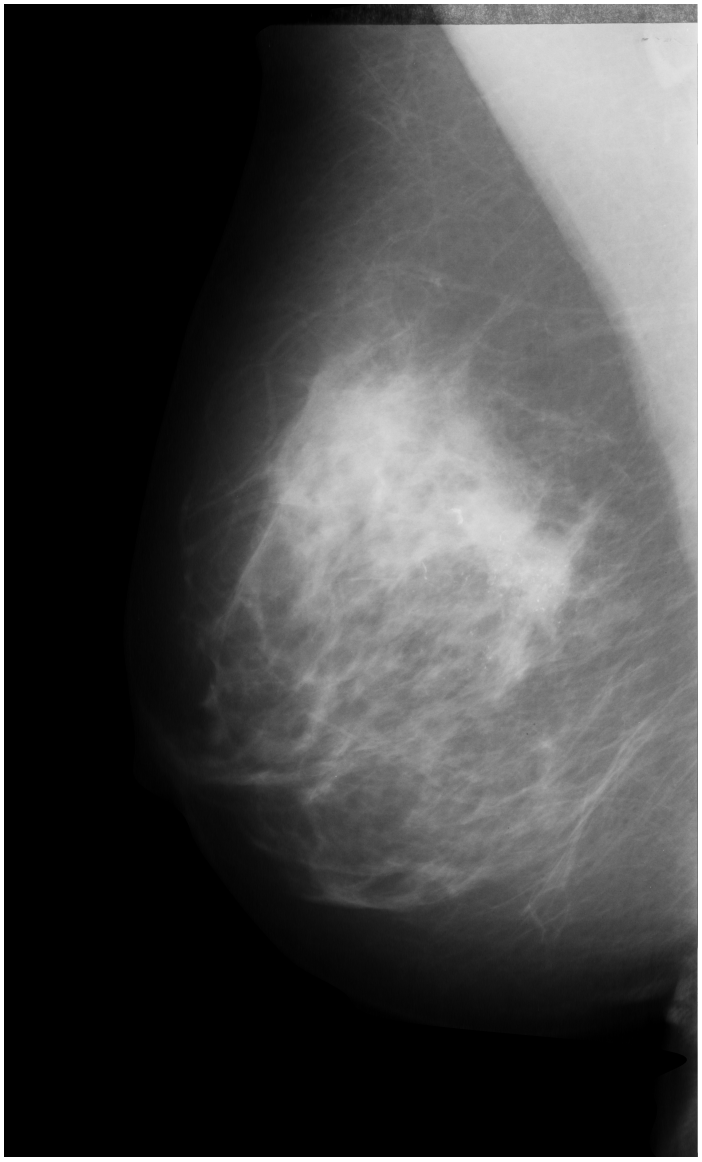


Figura 5.7: Mamografía digital correspondiente al ítem 209 del dataset.

5.2.2 Inferencia utilizando una ANN

El resultado de la inferencia que presenta la red neuronal artificial se puede apreciar en la figura 5.8 donde las zonas color magenta corresponden a la opción de rechazo, las verdes a la clase sana y las rojas a la clase con presencia de MCs.

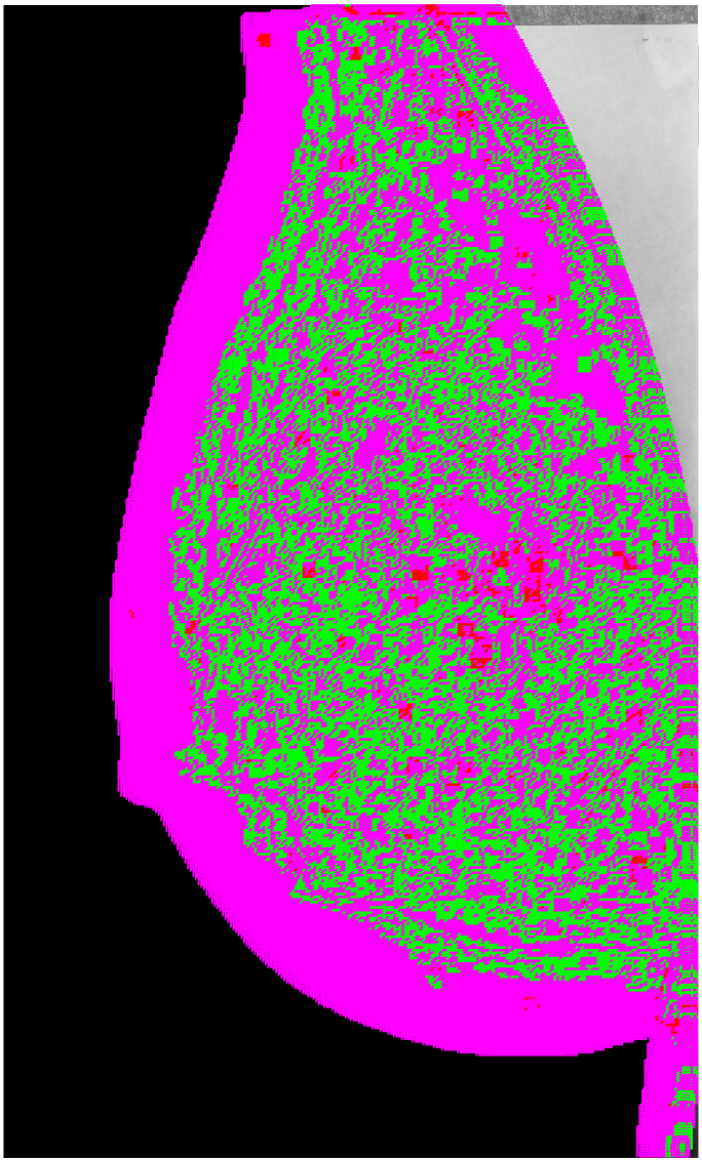


Figura 5.8: Inferencia con ANN, con recuadros de colores para cada clase. [Fuente: Elaboración propia]

Ahora, en la figura 5.9 se muestran solo las rojas pertenecientes a la clase con presencia de MCs sobre la imagen original.

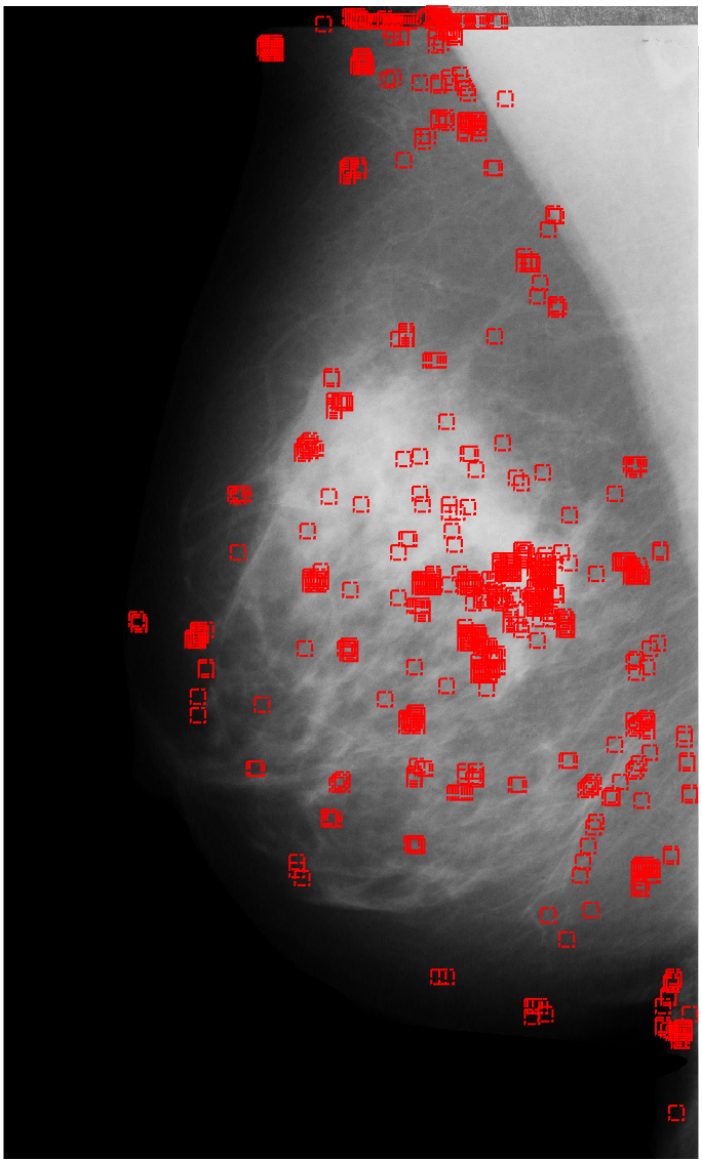


Figura 5.9: Inferencia con ANN, visualizando con recuadros de color rojo la clase MCs. [Fuente: Elaboración propia]

5.2.2.1 Post-validación nivel 1

El resultado de aplicar la post-validación nivel 1 se puede apreciar en la figura 5.10 donde las zonas color magenta corresponden a la opción de rechazo, las verdes a la clase sana y las rojas a la clase con presencia de MCs. Se puede apreciar que la opción de rechazo aumentó considerablemente y las zonas con MCs se vieron reducidas, sin embargo, estas regiones evidentemente corresponden a microcalcificaciones como se

observará más adelante.

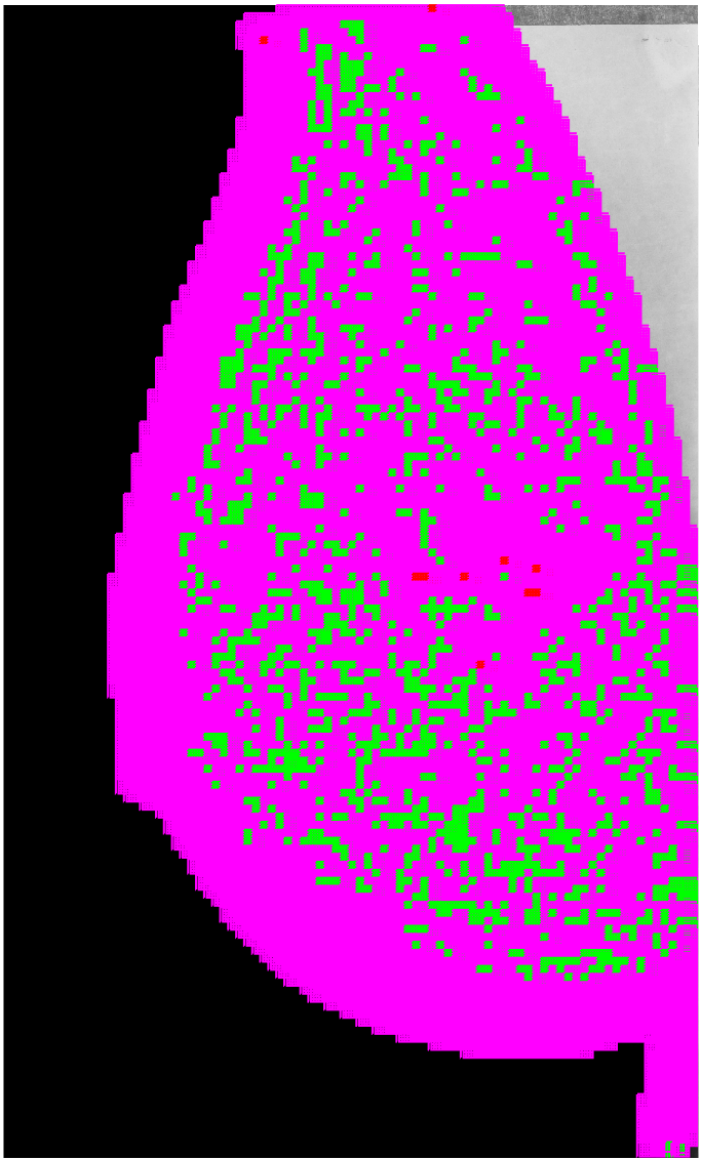


Figura 5.10: Imagen con recuadros de colores para cada clase con la post validación de nivel 1. [Fuente: Elaboración propia]

En la figura 5.11 se muestran solo las rojas después de realizar la post-validación de nivel 1.

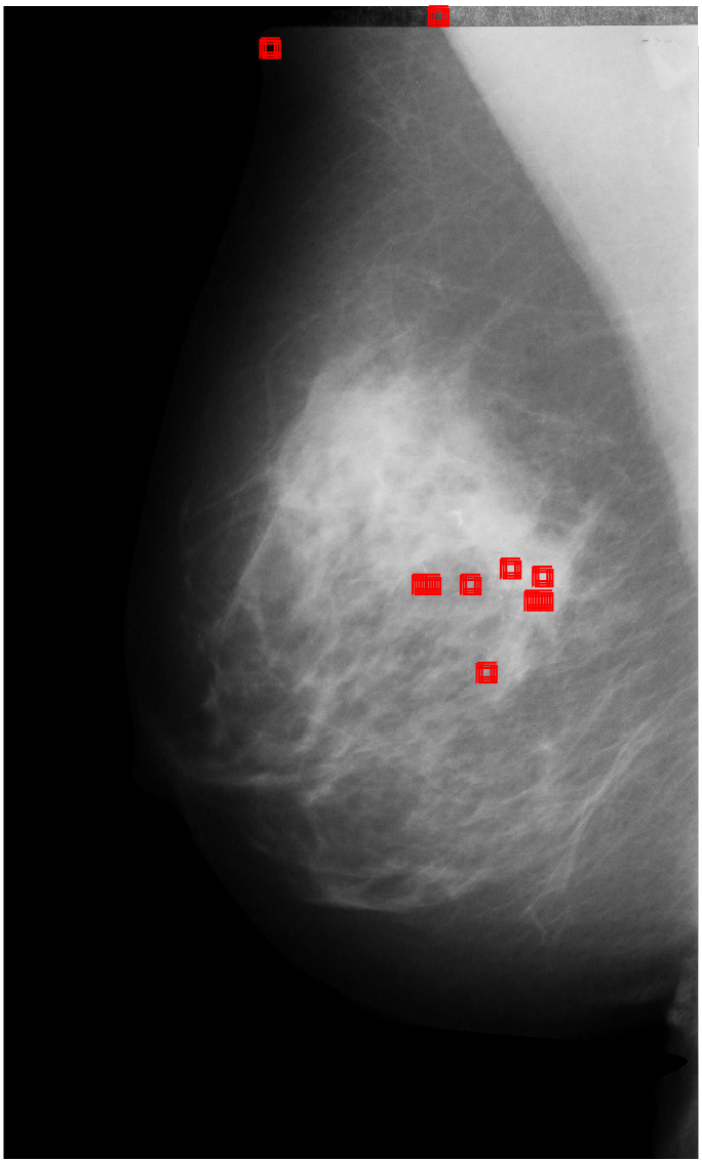


Figura 5.11: Imagen con recuadros de color rojo para la clase MCs únicamente con la post validación de nivel 1. [Fuente: Elaboración propia]

5.2.2.2 Post-validación nivel 2

El resultado de aplicar la post-validación nivel 1 se puede apreciar en la figura 5.12 donde las zonas color magenta corresponden a la opción de rechazo, las verdes a la clase sana y las rojas a la clase con presencia de MCs. Se puede apreciar que la opción de rechazo aumentó aún más con respecto a la post-validación nivel 1. En la figura 5.13 se muestran solo las rojas después de realizar la post-validación de nivel 2, se puede

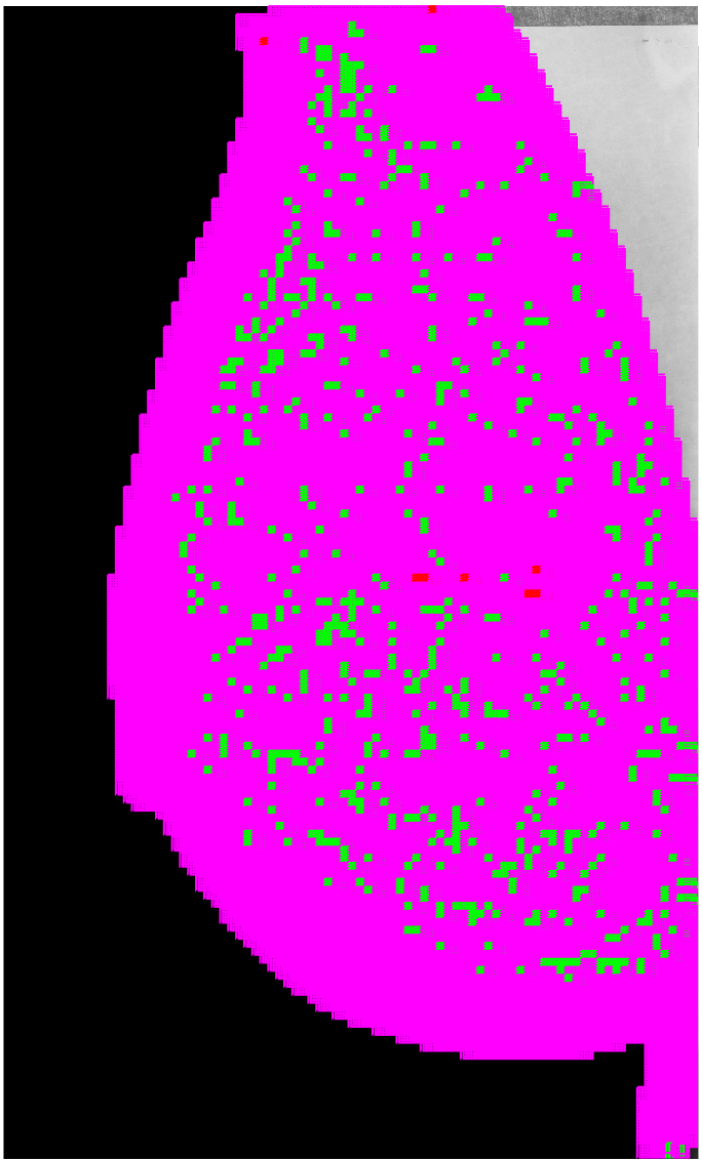


Figura 5.12: Imagen con recuadros de colores para cada clase con la post validación de nivel 2. [Fuente: Elaboración propia]

observar que se redujo el número de falsos positivos de la clase con MCs presentes en la imagen de inferencia.

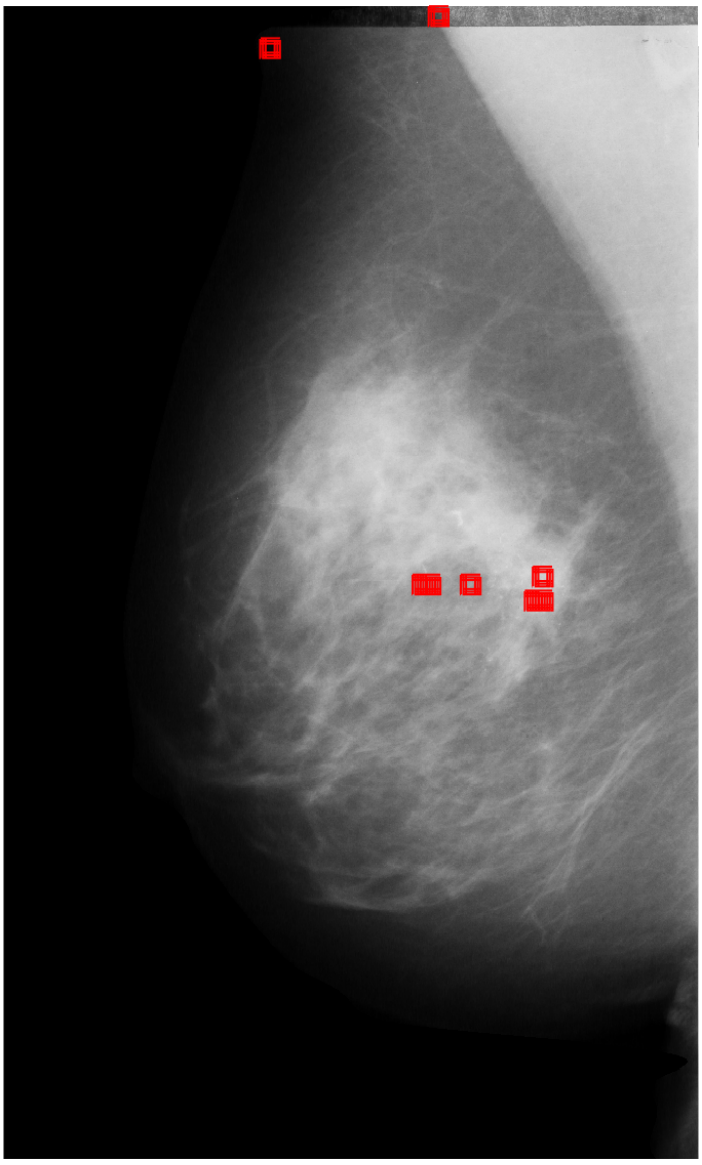


Figura 5.13: Imagen con recuadros de color rojo para la clase MCs únicamente con la post validación de nivel 2. [Fuente: Elaboración propia]

5.2.3 Inferencia utilizando una SVM

El resultado de la inferencia que presenta la máquina de soporte vectorial se puede apreciar en la figura 5.14 , donde se muestran solo las cajas rojas pertenecientes a la clase con presencia de MCs sobre la imagen original.

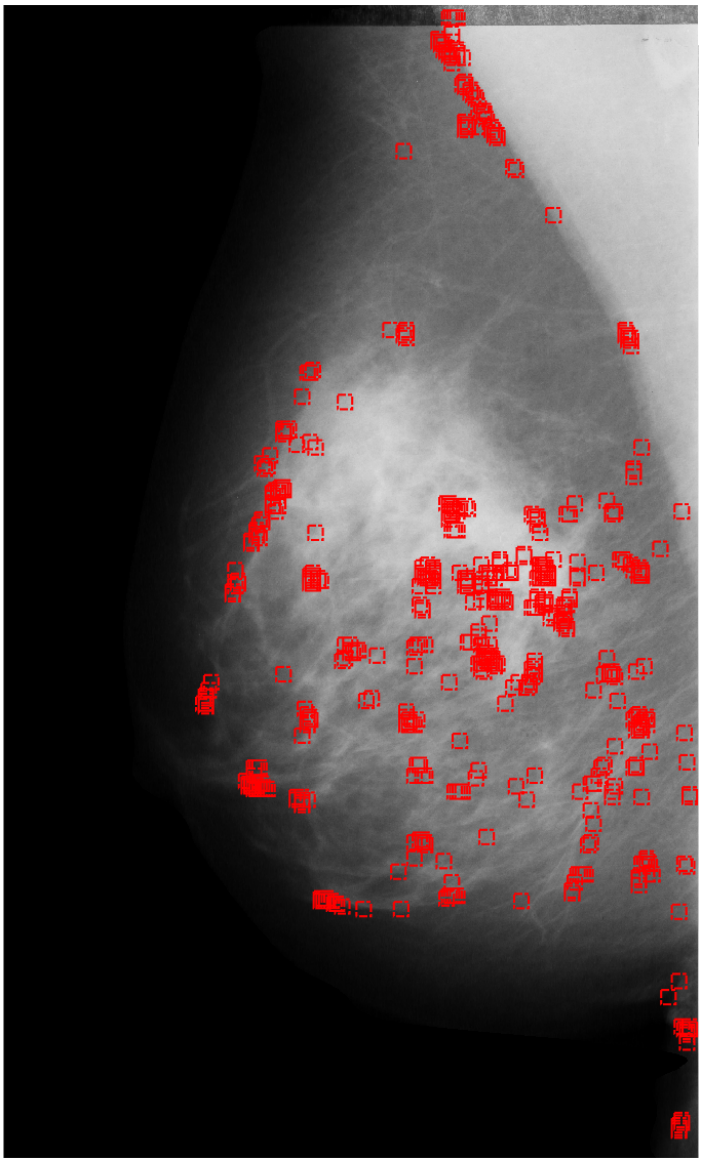


Figura 5.14: Inferencia con SVM, visualizando con recuadros de color rojo la clase MCs. [Fuente: Elaboración propia]

5.2.3.1 Post-validación nivel 1

El resultado de aplicar la post-validación nivel 1 se puede apreciar en la figura 5.15, donde se muestran solo las rojas después de realizar la post-validación de nivel 1.

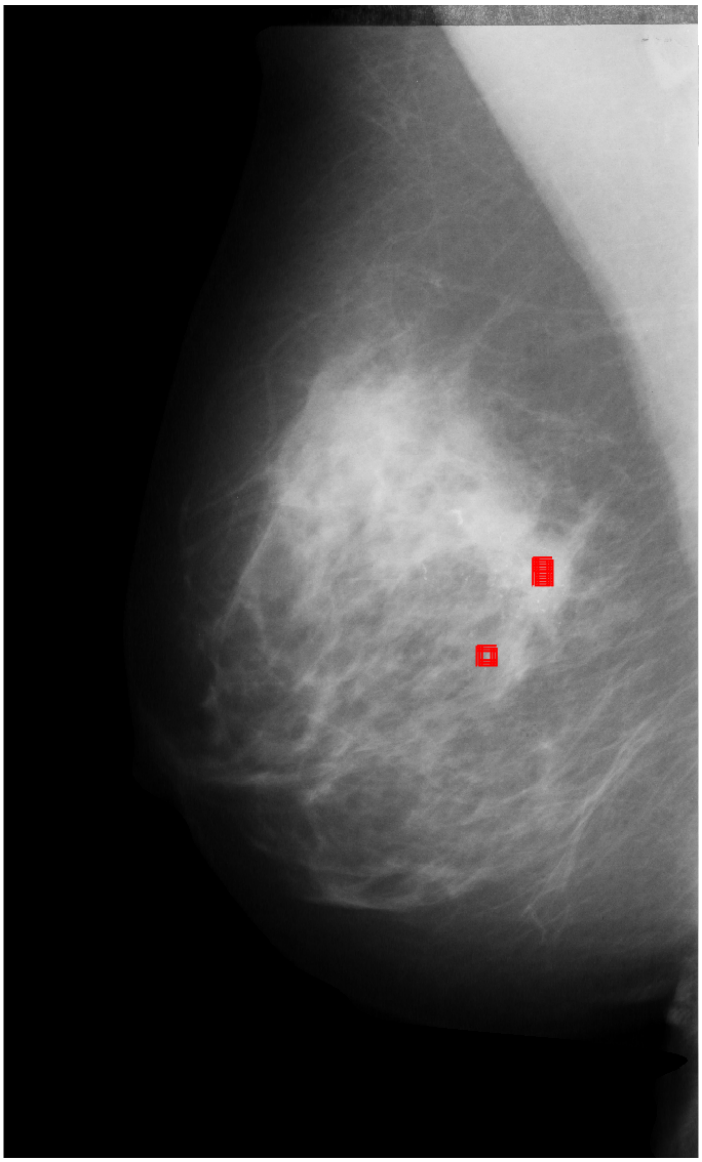


Figura 5.15: Inferencia con SVM, con recuadros de color rojo para la clase MCs únicamente con la post validación de nivel 1. [Fuente: Elaboración propia]

5.2.3.2 Post-validación nivel 2

El resultado de aplicar la post-validación nivel 1 se puede apreciar en la figura 5.16, donde se muestran solo las rojas después de realizar la post-validación de nivel 2, se puede observar que se redujo el número de falsos positivos de la clase con MCs presentes en la imagen de inferencia..

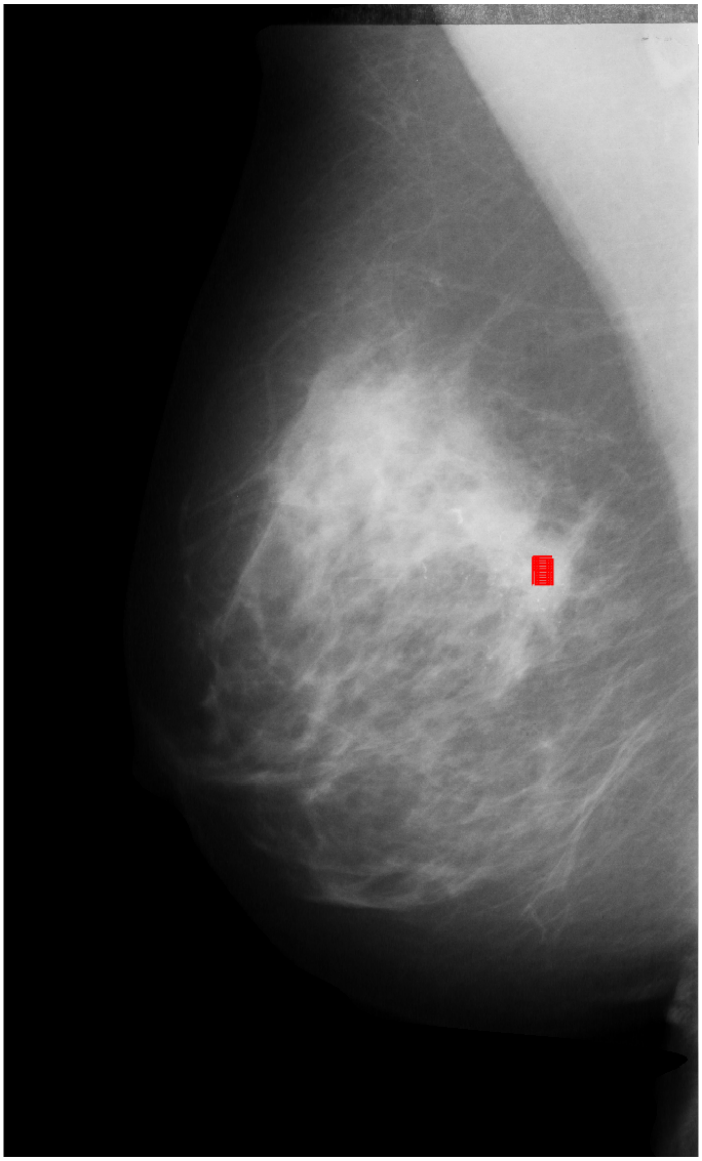


Figura 5.16: Inferencia con SVM, con recuadros de color rojo para la clase MCs únicamente con la post validación de nivel 2. [Fuente: Elaboración propia]

5.2.4 Imagen 223

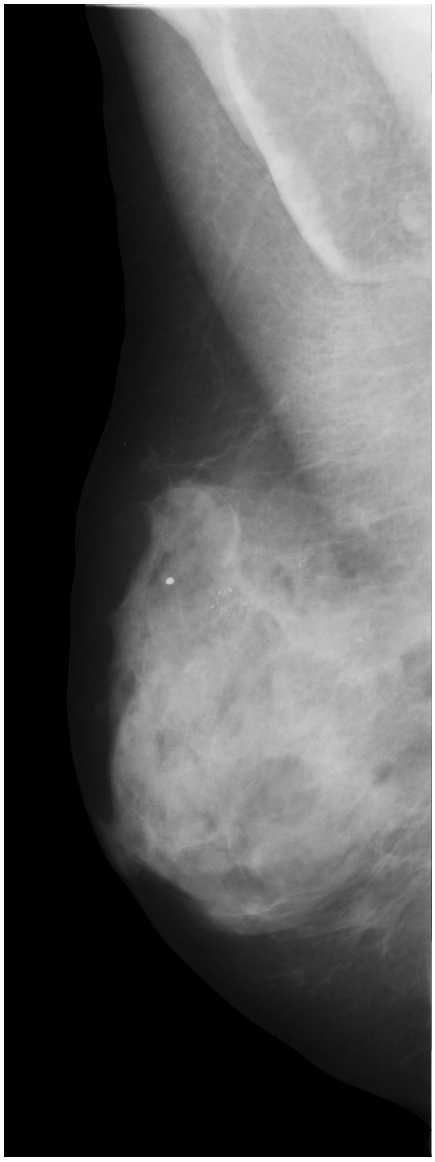


Figura 5.17: Imagen 223 del dataset. [Fuente: Elaboración propia]

5.2.5 Inferencia utilizando una ANN

El resultado de la inferencia que presenta la red neuronal se puede apreciar en la figura 5.18, donde se muestran solo las cajas rojas pertenecientes a la clase con presencia de MCs sobre la imagen original.

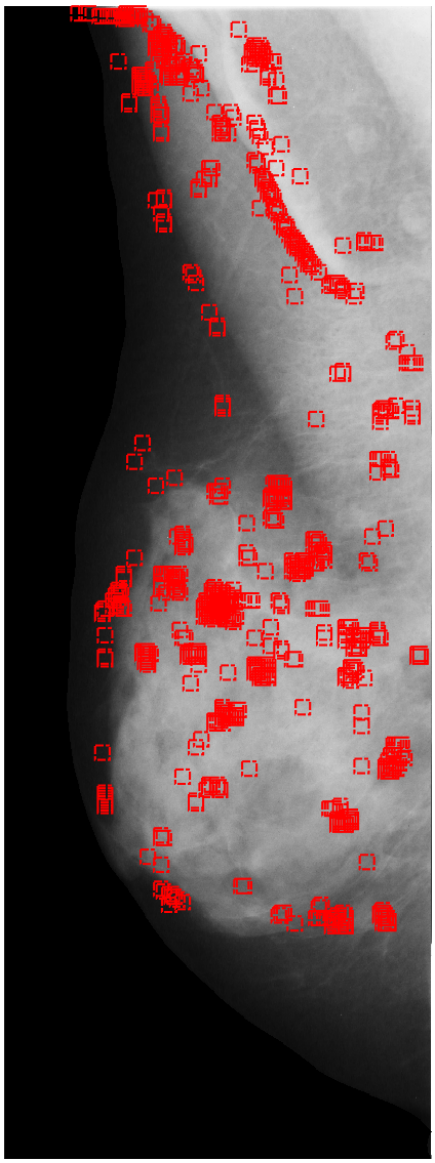


Figura 5.18: Inferencia con ANN, visualizando con recuadros de color rojo la clase MCs. [Fuente: Elaboración propia]

5.2.5.1 Post-validación nivel 1

El resultado de aplicar la post-validación nivel 1 se puede apreciar en la figura 5.19, donde se muestran solo las rojas después de realizar la post-validación de nivel 1.

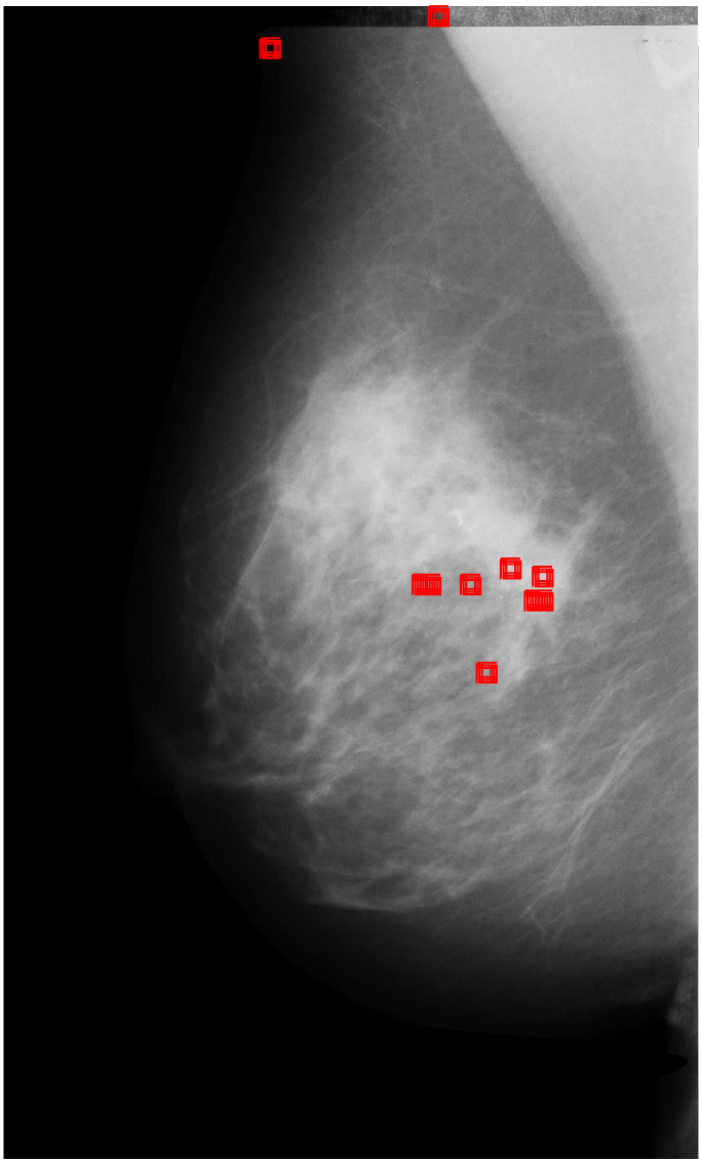


Figura 5.19: Inferencia con ANN, con recuadros de color rojo para la clase MCs únicamente con la post validación de nivel 1. [Fuente: Elaboración propia]

5.2.5.2 Post-validación nivel 2

El resultado de aplicar la post-validación nivel 1 se puede apreciar en la figura 5.20, donde se muestran solo las rojas después de realizar la post-validación de nivel 2, se puede observar que se redujo el número de falsos positivos de la clase con MCs presentes en la imagen de inferencia..

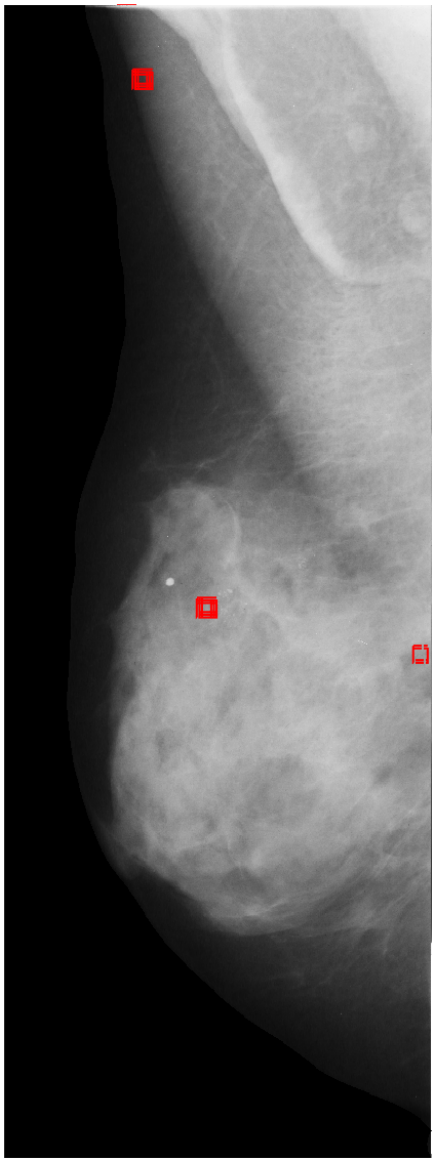


Figura 5.20: Inferencia con ANN, con recuadros de color rojo para la clase MCs únicamente con la post validación de nivel 2. [Fuente: Elaboración propia]

5.2.6 Inferencia utilizando una SVM

El resultado de la inferencia que presenta la máquina de soporte vectorial se puede apreciar en la figura 5.21 , donde se muestran solo las cajas rojas pertenecientes a la clase con presencia de MCs sobre la imagen original.

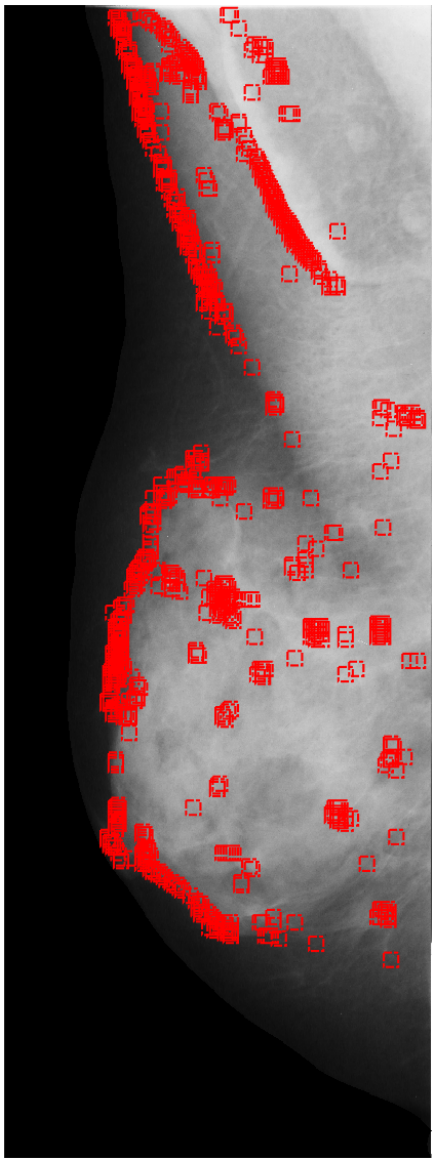


Figura 5.21: Inferencia con SVM, visualizando con recuadros de color rojo la clase MCs. [Fuente: Elaboración propia]

5.2.6.1 Post-validación nivel 1

El resultado de aplicar la post-validación nivel 1 se puede apreciar en la figura 5.22, donde se muestran solo las rojas después de realizar la post-validación de nivel 1.

Figura 5.22: Inferencia con SVM, con recuadros de color rojo para la clase MCs únicamente con la post validación de nivel 1. [Fuente: Elaboración propia]

5.2.6.2 Post-validación nivel 2

El resultado de aplicar la post-validación nivel 1 se puede apreciar en la figura 5.23, donde se muestran solo las rojas después de realizar la post-validación de nivel 2, se puede observar que se redujo el número de falsos positivos de la clase con MCs presentes en la imagen de inferencia..

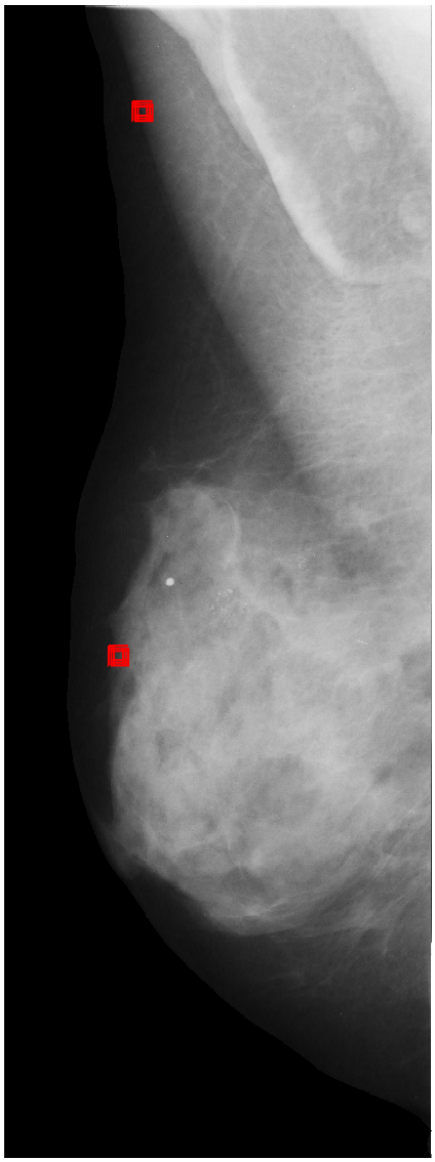


Figura 5.23: Inferencia con SVM, con recuadros de color rojo para la clase MCs únicamente con la post validación de nivel 2. [Fuente: Elaboración propia]

5.3 Evaluación de resultados

Con el fin de seleccionar el clasificador con el mejor desempeño para este problema en específico, se decidió tomar un *ground truth* conocido previamente y compararlo con las inferencias realizadas anteriormente, en esta sección se incluyen distintos modelos, así como clasificadores entrenados con los vectores ponderados con datos sin normalizar y con la tabla final de ponderaciones, esto con el fin de realizar una comparación objetiva

y seleccionar el mejor de todos los implementados.

5.3.1 Selección de métricas

En un principio se descartó la métrica de exactitud, debido a que la imagen con la que se realizará el testeo o prueba, no está balanceada para ambas clases (sana y presencia de MCs), es decir, no hay la misma cantidad de datos positivos y negativos, por ende, utilizar esta métrica sería incurrir en información no fiable. Por otro lado, se seleccionó la métrica de la precisión para las clases positivas y la métrica de sensibilidad; estas métricas se ajustan de mejor manera a las imágenes utilizadas y tienen en cuenta los falsos negativos y verdaderos positivos por encima de los falsos positivos, los cuales no son tan críticos como los falsos negativos (clasificación de tejido sano cuando en realidad hay presencia de MCs).

Cabe resaltar que la métrica de sensibilidad se alteró para que evaluara críticamente el problema específico, ya que el *ground truth* corresponde a una zona donde se encuentran MCs y no que toda la zona es perteneciente a la clase. Debido a esto, se decidió ponderar la sensibilidad de acuerdo a la ecuación 5.1.

$$s = \frac{0.9 \cdot TP}{0.9 \cdot TP + 0.1 \cdot FN} \tag{5.1}$$

Gracias a esta ponderación, se le recompensará más a la máquina por que acierte a la región donde se encuentren las MCs que por el porcentaje de la región que acierte.

5.3.2 Comparación del *ground truth* con la inferencia en la imagen 209

Para todas las comparaciones se utilizará la siguiente convención de colores:

- Verde: *Ground truth*.
- Magenta: Predicción.
- Blanco: Intersección entre el *Ground truth* y la Predicción.
- Negro: Zona sin predicción, *Ground truth* ni intersección entre estas.

5.3.2.1 SVM con vector de ponderaciones sin normalizar

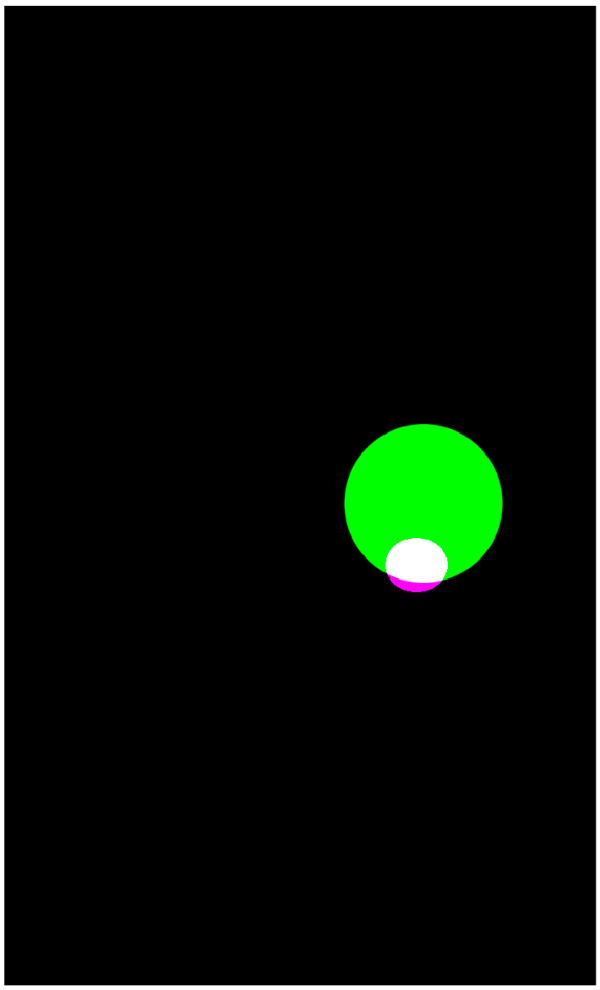


Figura 5.24: Comparación de la máquina de soporte vectorial con el vector de ponderaciones sin normalizar. [Fuente: elaboración propia]

La precisión resultante fue 0.844 y la sensibilidad 0.537.

5.3.2.2 SVM con vector de ponderaciones total

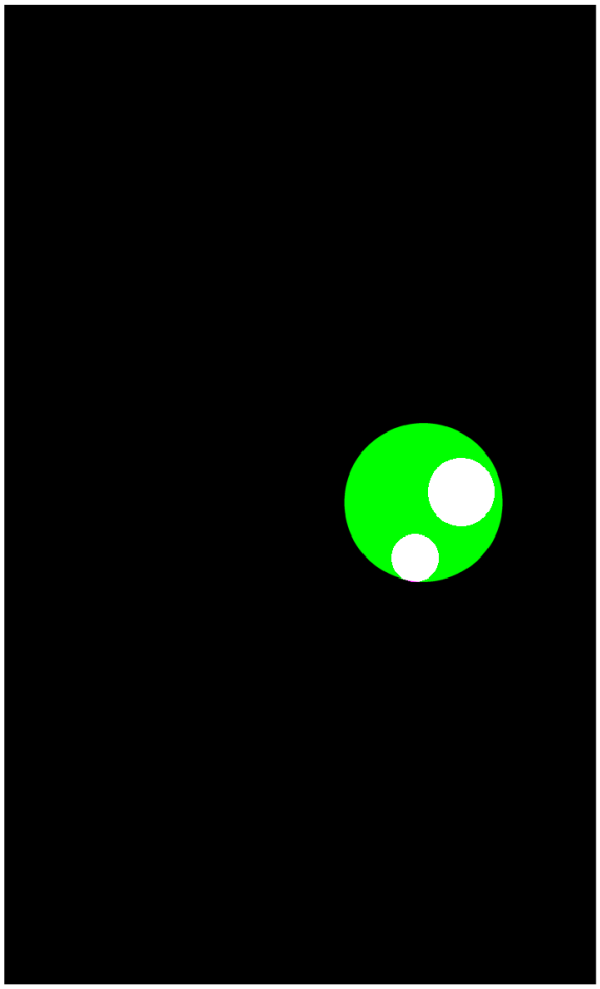


Figura 5.25: Comparación de la máquina de soporte vectorial con el vector de ponderaciones total. [Fuente: elaboración propia]

La precisión resultante fue 0.998 y la sensibilidad 0.770.

5.3.2.3 Red neuronal con vector de ponderaciones sin normalizar

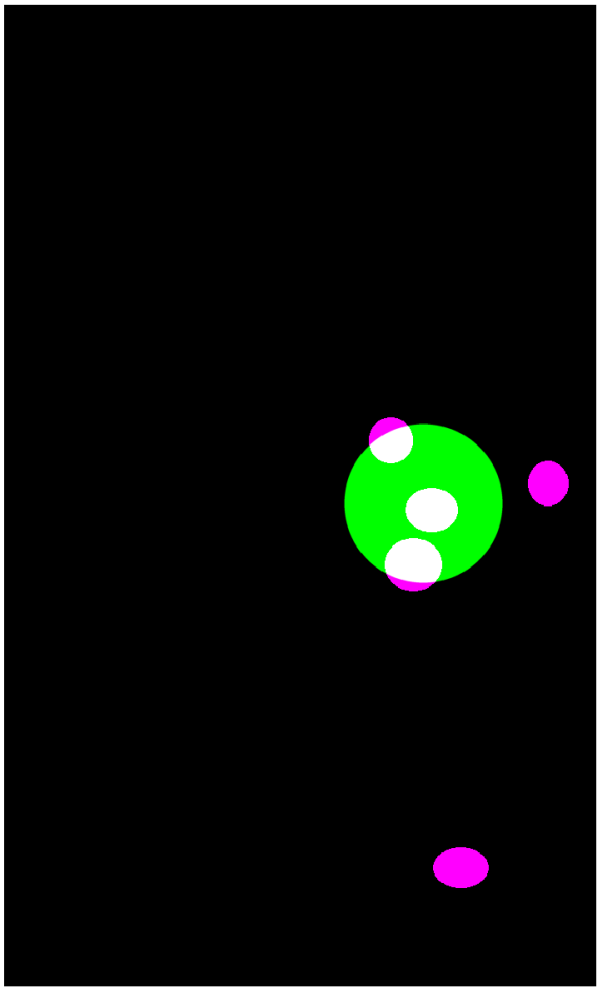


Figura 5.26: Comparación de la red neuronal con el vector de ponderaciones sin normalizar. [Fuente: elaboración propia]

La precisión resultante fue 0.555 y la sensibilidad 0.754.

5.3.2.4 Red neuronal con vector de ponderaciones total

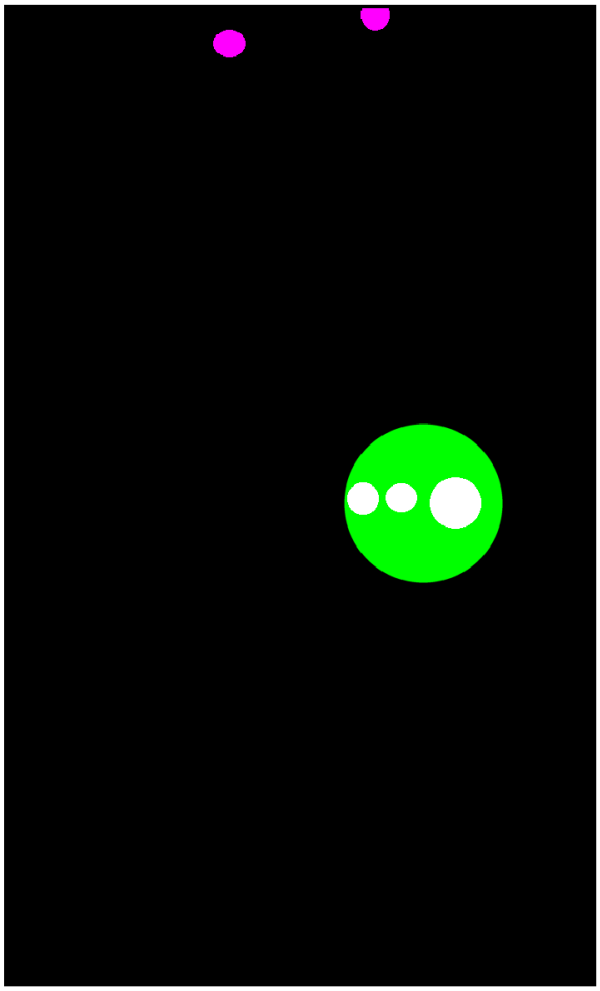


Figura 5.27: Comparación de la red neuronal con el vector de ponderaciones total. [Fuente: elaboración propia]

La precisión resultante fue 0.747 y la sensibilidad 0.668.

5.3.2.5 KNN con vector de ponderaciones total

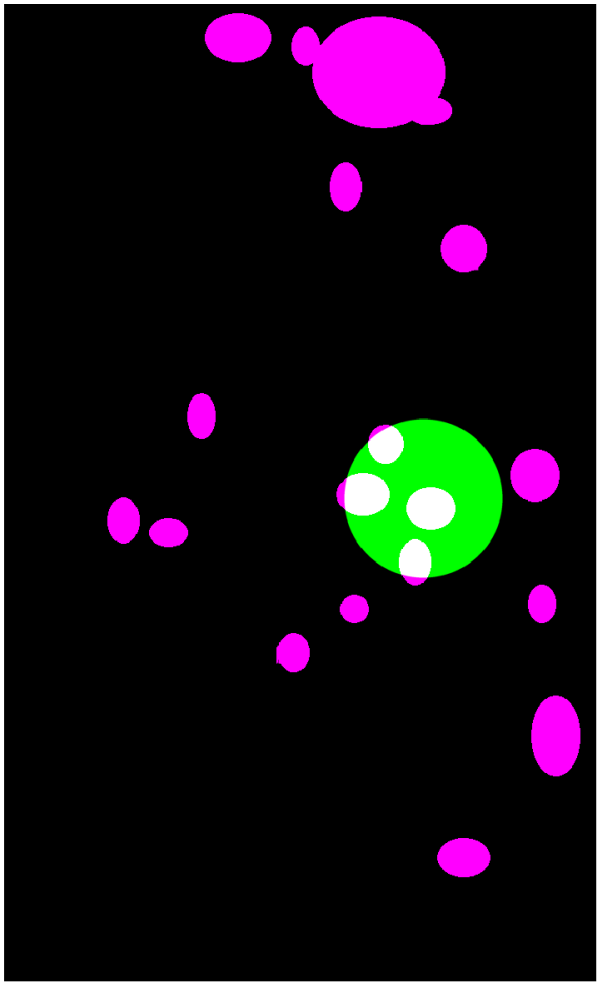


Figura 5.28: Comparación de KNN con el vector de ponderaciones total. [Fuente: elaboración propia]

La precisión resultante fue 0.142 y la sensibilidad 0.764.

5.3.2.6 Bayes con vector de ponderaciones total

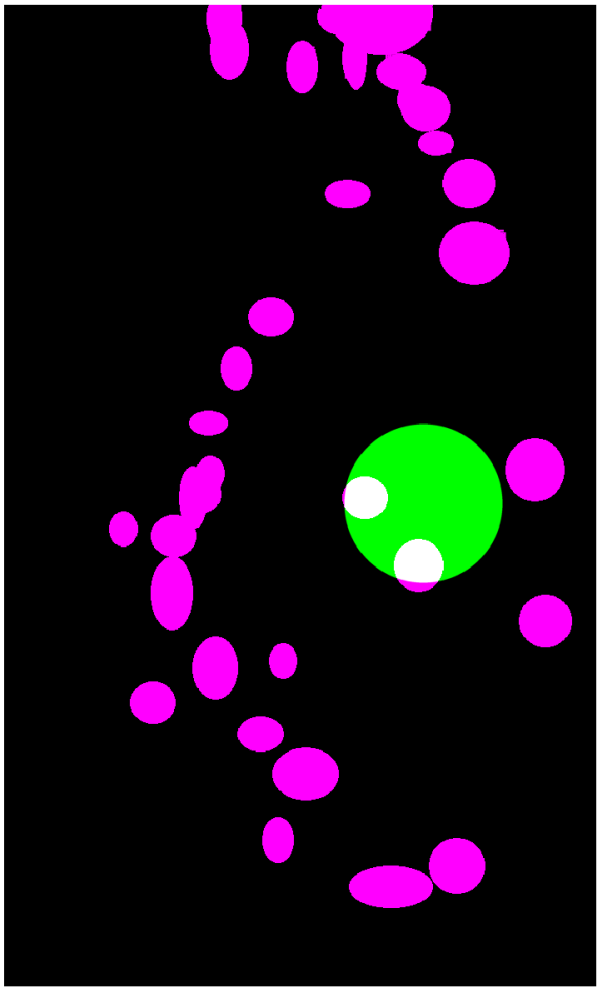


Figura 5.29: Comparación da bayes con el vector de ponderaciones total. [Fuente: elaboración propia]

La precisión resultante fue 0.060 y la sensibilidad 0.640.

5.3.3 Comparación del *ground truth* con la inferencia en la imagen 223

5.3.3.1 SVM con vector de ponderaciones total

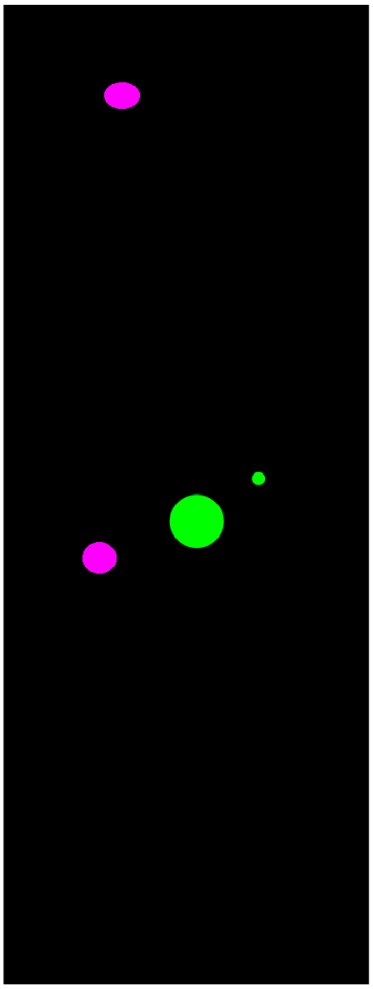


Figura 5.30: Comparación de la máquina de soporte vectorial con el vector de ponderaciones total. [Fuente: elaboración propia]

La precisión resultante fue 0 y la sensibilidad 0.

5.3.3.2 Red neuronal con vector de ponderaciones total

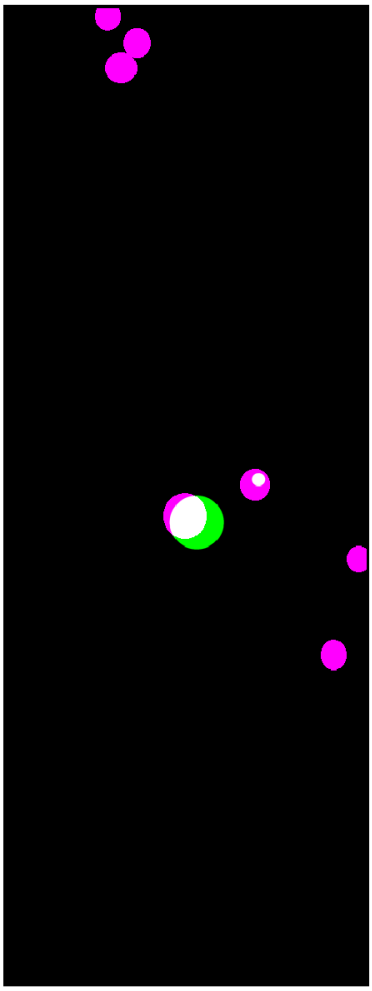


Figura 5.31: Comparación de la red neuronal con el vector de ponderaciones total. [Fuente: elaboración propia]

La precisión resultante fue 0.256 y la sensibilidad 0.921.

5.3.3.3 KNN con vector de ponderaciones total

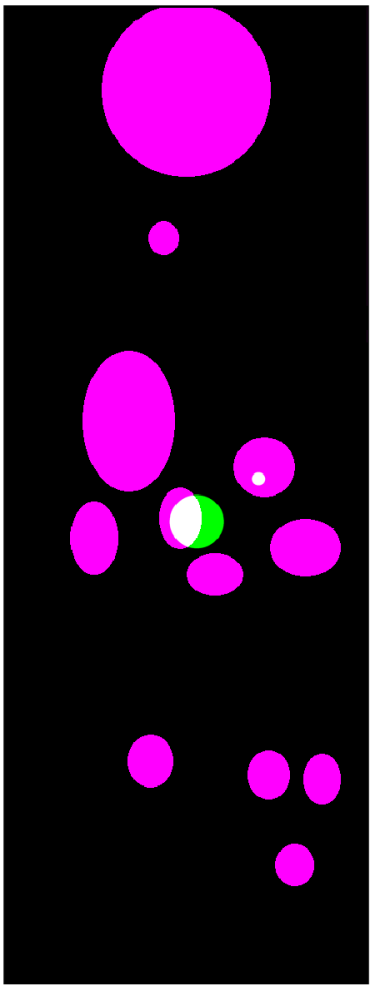


Figura 5.32: Comparación de KNN con el vector de ponderaciones total. [Fuente: elaboración propia]

La precisión resultante fue 0.025 y la sensibilidad 0.919.

5.4 Resumen de resultados

Después de realizar el cálculo de las métricas para la imagen N° 209, se procede a agruparlos en la tabla 5.5, en donde se pueden apreciar la precisión y la sensibilidad para cada caso de ponderación de características (total y sin normalizar) en las dos máquinas implementadas. Se puede apreciar que la ponderación total tuvo en general, mejores resultados que sin normalizar.

Imagen 209			
SVM			
Sin normalizar		**Total**	
p	**s**	**p**	**s**
0.844	0.537	0.998	0.770
RNA			
Sin normalizar		**Total**	
p	**s**	**p**	**s**
0.555	0.754	0.747	0.668

Tabla 5.5: Tabla de resultados para la imagen 209

Finalmente, en la tabla 5.6 se aprecian las métricas para las 2 imágenes (209 y 223) utilizando la ponderación total, ya que esta es la que mejor desempeño presentó en la métrica de mayor interés que es la sensibilidad. Cabe resaltar que la SVM no fue capaz de reconocer las MCs presentes en la imagen N° 223 mientras que las de RNA y K-NN lo hicieron de manera exitosa.

	Indices con ponderación total							
	SVM		**RNA**		**KNN**		**Bayes**	
	209	223	209	223	209	223	209	223
p	0.998	0	0.747	0.256	0.142	0.025	0.060	-
s	0.770	0	0.668	0.921	0.764	0.919	0.640	-

Tabla 5.6: Tabla de resultados para la ponderación total

6. Conclusiones

- **S**e desarrolló una metodología que permitió implementar un sistema de visión artificial para la detección de microcalcificaciones en mamografías. El desarrollo consta de etapas como preprocesamiento, extracción de características, selección de características, entrenamiento de cuatro tipos de clasificadores y detección de las microcalcificaciones sobre las mamografías.

- **S**e realizó una estrategia de extracción de características usando las reportadas en la literatura, con la cual se obtuvo inicialmente un vector de 149 características basado en descriptores estadísticos, de textura en el dominio espacial y textura en el dominio frecuencial. Posteriormente, se realizaron cuatro tipos de normalizaciones a la base de datos de entrenamiento con las cuales se aplicaron algoritmos de *ranking* para ponderar las características y seleccionar las más discriminantes para cada clase obteniendo un vector final de 15 características.

- **S**e implementó una etapa de preprocesamiento para la segmentación automática que separa el seno (región de interés para la inspección) del fondo de la imagen y el tejido pectoral usando técnicas de visión artificial para el realce del borde del seno mediante la transformación gamma, segmentación del seno y el fondo mediante la obtención del umbral por histograma del gradiente y supresión de artefactos presentes en la mamografía mediante la técnica de elementos conectados, permitiendo realizar una etapa de inferencia de los clasificadores entrenados en menor tiempo, consumiendo menos espacio en memoria y realizando la inspección en la región de interés.

- **S**e desarrollaron cuatro algoritmos de aprendizaje utilizando las máquinas RNA tipo MLP, SVM, K-NN y NB, para detectar tres clases: región con presencia de microcalcificaciones, región sana y una opción de rechazo. Se utilizó la métrica de exactitud para medir el desempeño de los clasificadores, la RNA obtuvo 97.54 %, la SVM 95.68 %, el K-NN 95.68 % y el NB 80.25 %, evaluando el dataset de validación. Finalmente, se realizó un análisis de varianza ANOVA con el fin de seleccionar el clasificador que presentó la distribución de predicciones

más diferenciadora, además del desempeñó obtenido en la etapa de detección e inferencia sobre las imágenes.

- **S**e determinaron alcances y limitaciones del sistema implementado donde se encontró que el uso del algoritmo de ventanas deslizantes para el proceso de inferencia presenta tiempos de cómputo muy grandes, si embargo, la posibilidad de parametrizar el tamaño de la ventana de inspección permitió la detección de microcalcificaciones de orden micrométrico incluso en regiones donde la visión humana no puede detectarlos con facilidad.

- **E**n el desarrollo de esta investigación se pudo corroborar que las características de textura son de vital importancia en la detección de microcalcificaciones en mamografías digitales, dado que en condiciones específicas de uniformidad sobre la región de inspección, se puede descartar la aparición de una microcalcificación directamente, concluyendo que la región puede ser sana o tratarse de otra patología radiológica que ocupa un mayor tamaño que la ventana de inspección, posibilitando que el sistema reduzca los tiempos de inspección sobre estas regiones y así identificar otra región de interés particular para el médico radiólogo.

- **L**a presencia de un tejido mamario denso sobre las mamografías digitales es un factor que dificulta la detección de microcalcificaciones, debido a que este se presenta en la imagen como una variación de intensidades en los niveles de gris que puede generar un alto índice de falsos positivos en el momento de la inferencia.

- **L**a segmentación del tejido pectoral en mamografías digitales es un proceso comúnmente realizado en los sistemas de detección de patologías relacionadas con el cáncer de mama, puesto que generalmente, en esta zona no hay presencia de indicadores relacionados con éste cáncer. Por lo anterior, aunque no es necesario, es importante su eliminación para evitar elevar el proceso de cómputo en la detección de alguna patología relacionada con el cáncer de mama.

- **L**a selección del tamaño de la ventana en el método de deslizamiento de ventanas es de vital importancia para un adecuado funcionamiento de preprocesamiento y extracción de características.

- **E**l análisis de varianza ANOVA es una herramienta que permite, entre otras cosas, medir qué tan robusto es un modelo con respecto a otro. Gracias a la implementación de este análisis en el proyecto se observó mediante diagramas de violín, la distribución de probabilidad de los clasificadores, esto permitió seleccionar la máquina que mejor discriminara entre las dos clases.

- **S**e seleccionó las métricas que miden críticamente el desempeño de los clasificadores, teniendo en cuenta su aplicación real en el campo de la medicina, en donde los FN (clasificación de tejido sano cuando en realidad hay presencia de MCs) son inaceptables.

- **S**e seleccionó y modificó la métrica de sensibilidad debido al desbalance natural presente en el *ground truth* proporcionado por el experto en el dataset MIAS.

- **L**a etiqueta proporcionada es una coordenada (x, y) y un radio r, esto da como resultado un círculo dentro del cual existen una o más MCs, sin embargo, la diferencia entre píxeles sanos frente a los que cuentan con presencia de MCs es de varios órdenes de magnitud. Debido a esto, métricas que castiguen los FN, como lo es la sensibilidad, van a dar como resultado valores bajos debido a la cantidad de píxeles sanos dentro del círculo proporcionado por el experto como clase con presencia de MCs.

- **E**l ajuste ponderado realizado a la métrica, obtuvo valores de sensibilidad cercanos al 70 %, sin embargo, para obtener un valor de esta métrica que describa de manera adecuada el desempeño real del clasificador, se requiere de un experto en radiología que etiquete píxel por píxel cada microcalcificación individualmente, así, se obtendrá un *ground truth* preciso de la zona con presencia de MCs que se quiere detectar.

- **S**e seleccionó la métrica de precisión para la clase con presencia de MCs, con el objetivo de medir el desempeño de clasificación netamente dentro del círculo de *ground truth* proporcionado.

7. Trabajos futuros

- Contactar a un experto en radiología con el objetivo de generar etiquetas específicas para las microcalcificaciones, esto permitirá una mayor confianza en la generación de los *datasets* de entrenamiento, validación y prueba, además de generar un *ground truth* más certero para la etapa de la evaluación de las métricas.
- Realizar la implementación para la detección de microcalcificaciones en mamografías digitales utilizando métodos de aprendizaje de máquina profundo para validar el alcance y desempeño de este tipo de arquitecturas con la realizada en este trabajo de investigación basado en maquinas de aprendizaje superficial.
- Implementar un desarrollo de visión por computador el cual incluya el uso de la característica llamada Unidades de Hounsfield (UH) que se encuentra presente en tomografías computarizadas, con la cual puede describir los diferentes niveles de radiodensidad de los tejidos humanos.
- Evuluar el desempeño de imagenes mamográficas tridimensionales para ser procesadas mediante técnicas de visión artificial y encontrar características discriminantes en la detección de microcalcificaciones.

Referencias

[A. Jáuregui et al., 1994] A. Jáuregui et al. (1994). Microcalcificaciones y cáncer de mama. https://www.sespm.es/wp-content/uploads/revista/1994_7_3/2.pdf. En línea; accedido 12 Dicembre 2018.

[Academia Nacional de Medicina de Colombia, 2016] Academia Nacional de Medicina de Colombia (2016). ¿A que edad iniciar la toma rutinaria? https://anmdecolombia.org.co/mamografias-a-que-edad-iniciar-la-toma-rutinaria/. En línea; accedido 13 Septiembre 2018.

[American Cancer Society, 2017] American Cancer Society (2017). Tratamiento del carcinoma ductal in situ. https://www.cancer.org/es/cancer/cancer-de-seno/tratamiento/tratamiento-del-cancer-del-seno-segun-su-etapa/tratamiento-del-carcinoma-ductal-in-situ.html. En línea; accedido 21 Septiembre 2018.

[Arancibia et al., 2016] Arancibia, P., Estrada, T. T., López, A., Díaz, M. L., and Sáez, C. (2016). Breast calcifications: Description and classification according to BI-RADS 5th edition. *Revista Chilena de Radiología - ScienceDirect*, 22:80–91.

[Basile et al., 2019] Basile, T., Fanizzi, A., Losurdo, L., Bellotti, R., Bottigli, U., Dentamaro, R., Didonna, V., Fausto, A., Massafra, R., M, M., Tamborra, P., Tangaro, S., and Forgia, D. L. (2019). Microcalcification detection in full-field digital mammograms: A fully automated computer-aided system. *European Journal of Medical Physics - ElSevier*, 64:1–9.

[BreastCancerOrganization, 2018] BreastCancerOrganization (2018). ¿Qué es el cáncer de mama? https://www.breastcancer.org/es/sintomas/cancer_de_mama/que_es_cancer_mama. En línea; accedido 11 Dicembre 2018.

[Costa et al., 2012] Costa, A. F., Humpire-Mamani, G., and Traina, A. J. M. (2012). An efficient algorithm for fractal analysis of textures. In *2012 25th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 39–46.

[Erickson et al., 2017] Erickson, B. J., Korfiatis, P., Akkus, Z., and Kline, T. L. (2017). Machine Learning for Medcial Imaging. *RadioGraphics*, 37 No.2:505–515.

[Esteva et al., 2019] Esteva, A., Robicquet, A., Ramsundar, B., and Kuleshov, V. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25:24–29.

[Fadil et al., 2019] Fadil, R., Jackson, A., Majd, B. A. E., Ghazi, H. E., and Kaabouch, N. (2019). Segmentation of microcalcifications in mammograms: A comparative study. In *International Conference on Electro-Information Technology EIT 2019 - IEEE Xplore*.

[Fadil et al., 2020] Fadil, R., Jackson, A., Majd, B. A. E., Ghazi, H. E., and Kaabouch, N. (2020). Classification of microcalcifications in mammograms using 2d discrete wavelet transform and random forest. In *International Conference on Electro-Information Technology EIT 2020 - IEEE Xplore*.

[Frangi et al., 1998] Frangi, A. F., Niessen, W. J., Vincken, K. L., and Viergever, M. A. (1998). Multiscale vessel enhancement filtering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention-1998. MICCAI - Springer*.

[Ge et al., 2006] Ge, J., Sahiner, B., Hadjiiski, L. M., Chan, H.-P., Wei, J., Helvie, M. A., and Zhou, C. (2006). Computer aided detection of clusters of microcalcifications on full field digital mammograms. *Medical Physics*, 33:2975–2988.

[Guo et al., 2010] Guo, Z., Zhang, L., and Zhang, D. (2010). A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing*, 19(6):1657–1663.

[Haralick et al., 1973] Haralick, R. M., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621.

[Henrot et al., 2014] Henrot, P., Leroux, A., Barlier, C., and Génin, P. (2014). Breast microcalcifications: The lesions in anatomical pathology. *Diagnostic and Interventional Imaging - ScienceDirect*, 95:141–152.

[Jundong Li et al., 2017] Jundong Li, K. C., Wang, S., Morstatter, F., Trevino, R. P., and J. Tang, H. L. (2017). Feature Selection: A Data Perspective. *Association for Computing Machinery*, 50-Article 94:94–45.

[K. Pesce et al., 2012] K. Pesce et al. (2012). Eficacia de la mamografía como método de screening para el diagnóstico del cáncer de mama. `https://www.hospitalitaliano.org.ar/multimedia/archivos/servicios_attachs/8277.pdf`. En línea; accedido 18 Diciembre 2018.

[Kumari et al., 2014] Kumari, A., Thomas, P. J., and Sahoo, S. (2014). Single image fog removal using gamma transformation and median filtering. In *2014 Annual IEEE India Conference (INDICON)*, pages 1–5.

[Lee et al., 2017] Lee, A. Y., Wisner, D. J., Aminololama-Shakeri, S., Arasu, V. A., Feig, S. A., Hargreaves, J., Ojeda-Fournier, H., Bassett, L. W., Wells, C. J., Guzman, J. D., l. Flowers, C., Campbell, J. E., Elson, S. L., Retallack, H., and Joe, B. N. (2017). Inter-reader Variability in the Use of BI-RADS Descriptors for Suspicious Finding on Diagnostic Mammography: A Multi-institution Study of 10 Academic Radiologists. *Academic Radiology - ScienceDirect*, 24:60–66.

[Liu and Yu, 2009] Liu, D. and Yu, J. (2009). Otsu method and k-means. In *2009 Ninth International Conference on Hybrid Intelligent Systems*, volume 1, pages 344–349.

[Löfstedt et al., 2019] Löfstedt, T., Brynolfsson, P., Asklund, T., Nyholm, T., and Garpebring, A. (2019). Gray-level invariant haralick texture features. *PLOS ONE*, 14:e0212110.

[Ma et al., 2019] Ma, J., Ma, Y., and Li, C. (2019). Infrared and visible image fusion methods and applications: A survey. *Information Fusion - ScienceDirect*, 45:153–178.

[Moreira et al., 2012] Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., and Cardoso, J. S. (2012). Inbreast: Toward a full-field digital mammographic database. *Academic Radiology*, 19(2):236–248.

[Nielsen et al., 2014] Nielsen, M., Vachon, C. M., Scott, C. G., Chernogg, K., Karemore, G., Karssemeijer, N., Lillholms, M., and Karsdal, M. A. (2014). Mammographic texture resemblance generalizes as an independent risk factor for breast cancer. *Breast Cancer Research - Part of Springer Nature*, 16(2):R37.

[Padrón-Godínez et al., 2008] Padrón-Godínez, A., Lee, M., Becerra, A., Prieto Meléndez, R., and Robles, J. (2008). Ocultamiento de datos en imágenes digitales mediante bpcs. In *Universidad autónoma de México*.

[Papadopoulos et al., 2002] Papadopoulos, A., Fotiadis, D. I., and Likas, A. (2002). An automatic microcalcification detection system based on a hybrid neural network classifier. *Artificial Intelligence in Medicine - ElSevier*, pages 149–167.

[Pertuz et al., 2019] Pertuz, S., Torres, G., Tamimi, R., and Kämäräinen, J.-K. (2019). Open framework for mammography-based breast cancer risk assessment. In *IEEE-EMBS Int. Conf. on Biomedical and Healtch Informatics (BHI)*.

[Rincón et al., 2018] Rincón, J., Castro, A., Narváez, F., and Díaz, G. (2018). Machine learning methods for classifying mammographic regions using the wavelet transform and radiomic texture features. In *International Conference on Technology Trends CITT 2018. Communications in Computer and Information Science - Springer*.

[Salas, 2004] Salas, R. (2004). Redes neuronales artificiales. *Universidad de Valparaıso. Departamento de Computación*, 1.

[Sartori et al., 2015] Sartori, P., Rozowykniat, M., Siviero, L., Barba, G., Peña, A., Mayol, N., Acosta, D., Castro, J., and Ortiz, A. (2015). Artefactos y artificios frecuentes en tomografía computada y resonancia magnética. *Revista Argentina de Radiología*, 79(4):192–204.

[Shin et al., 2015] Shin, S. Y., Lee, S., Yun, D., Jung, H. Y., Heo, Y. S., Kim, S. M., and Lee, K. M. (2015). A Novel Cascade Classifier for Automatic Microcalcification Detection. *PLoS One*, pages 1–22.

[Sickles, 1986] Sickles, E. A. (1986). Breast Calcifications: Mammographic Evaluation. *Radiology - Radiological Society of North America (RSNA)*, 60:289–293.

[Szeliski, 2010] Szeliski, R. (2010). *Computer vision: algorithms and applications*. Springer Science & Business Media.

[Torres and Pertuz, 2016] Torres, G. and Pertuz, S. (2016). Automatic detection of the retroareolar region in x-ray mammography images. In *VII Latin American Congress on Biomedical Engineering CLAIB 2016*.

[University of Aveiro, 2008] University of Aveiro (2008). Breast Cancer Digital Repository. `https://bcdr.eu/`. En línea.

[Vargas et al., 2014] Vargas, H. D., Orozco, A., and Alvarez, M. A. (2014). Automatic Recognition of Microcalcifications in Mammography Images through Fractal Texture Analysis. *Advances in Visual Computing Part II. 10th International Symposium, ISVC 2014 - SpringerLink*, 8888:841–850.

[Zheng et al., 2015] Zheng, Y., Keller, B. M., Ray, S., Wang, Y., Conant, E. F., Gee, J. C., and Kontos, D. (2015). Parenchymal texture analysis in digital mammography: A fully automated pipeline for breast cancer risk assessment. *Medical Physics*, 42(7):4149–4160.

Apéndice A

Tabla completa de ponderación de características por algoritmo

	N° de carácterística				
Rank	**Sin normalizar**	**De 0 a 1**	**De -1 a 1**	**Estandarización**	**Soft-max**
1	126	126	126	3	3
2	3	138	138	126	126
3	140	127	127	127	127
4	125	125	3	138	140
5	149	140	140	140	138
6	138	3	125	124	121
7	4	149	124	125	125
8	124	121	149	121	124
9	143	124	121	149	122
10	147	4	4	4	4
11	139	30	102	139	149
12	137	102	139	137	46
13	127	139	147	128	102
14	121	32	137	147	147
15	2	137	30	30	60
16	44	46	128	32	32
17	122	128	32	102	128
18	141	88	88	46	137
19	30	100	86	141	2
20	35	60	60	143	44
21	135	147	100	122	117
22	46	86	44	2	35
23	123	44	58	100	139
24	134	91	77	117	30
25	58	116	135	123	58
26	32	58	91	44	100

27	102	59	116	88	88
28	60	77	18	134	91
29	100	21	46	135	86
30	136	74	74	21	49
31	49	117	117	86	141
32	91	114	114	18	120
33	148	135	123	91	21
34	21	101	134	35	77
35	18	87	87	60	116
36	128	123	2	59	114
37	86	134	143	116	87
38	72	35	21	77	135
39	74	143	49	114	134
40	88	18	59	58	143
41	77	31	31	74	123
42	109	49	72	31	105
43	16	45	63	72	18
44	131	105	136	87	74
45	87	115	101	132	132
46	7	2	141	115	72
47	45	141	105	136	59
48	59	72	122	45	101
49	101	136	35	131	31
50	132	132	132	148	45
51	31	63	7	101	73
52	63	148	16	109	63
53	114	16	115	63	16
54	81	122	148	120	131
55	116	7	131	105	115
56	120	73	45	16	7
57	118	131	73	49	17
58	105	17	109	17	129
59	73	109	17	7	136

60	11	106	106	81	148
61	25	120	120	73	146
62	17	81	78	106	109
63	67	129	81	25	144
64	39	10	64	11	106
65	115	78	10	118	118
66	129	64	11	39	64
67	29	1	25	129	81
68	117	108	92	67	10
69	53	38	39	78	78
70	95	11	50	64	66
71	111	25	38	10	11
72	57	119	22	29	38
73	119	39	129	95	50
74	69	47	67	53	25
75	145	50	52	66	22
76	27	52	8	38	1
77	47	92	53	92	39
78	20	22	145	22	92
79	34	145	47	50	29
80	12	66	80	52	52
81	97	67	95	8	80
82	55	61	36	80	119
83	65	118	1	36	67
84	83	8	66	119	8
85	82	53	118	108	108
86	42	95	108	1	53
87	56	80	24	24	23
88	142	36	61	145	145
89	54	24	119	23	36
90	110	33	33	20	24
91	146	29	142	9	95
92	61	19	130	94	47

93	26	12	103	12	61
94	43	142	19	47	20
95	96	146	29	43	133
96	130	133	5	27	33
97	133	5	69	142	69
98	9	23	146	56	82
99	40	130	82	133	142
100	68	69	111	14	9
101	23	103	133	33	94
102	103	111	12	40	19
103	106	75	110	57	28
104	33	94	40	65	12
105	113	82	94	97	75
106	85	20	14	130	130
107	19	68	89	82	40
108	48	40	68	146	26
109	28	89	43	69	110
110	64	110	27	19	103
111	41	27	75	68	27
112	71	26	23	15	54
113	78	54	96	26	57
114	15	96	54	61	111
115	75	97	65	28	96
116	99	9	9	111	56
117	5	144	26	110	65
118	13	65	97	144	55
119	89	57	20	34	85
120	50	56	55	55	5
121	22	55	83	83	15
122	92	83	34	89	83
123	1	34	57	42	14
124	8	14	42	54	48
125	10	42	41	103	97

126	107	15	56	96	34
127	38	28	144	48	42
128	36	85	15	6	68
129	51	71	48	71	89
130	66	41	28	5	113
131	37	13	51	75	6
132	144	48	6	85	98
133	70	43	71	41	41
134	52	51	85	51	51
135	62	99	70	70	13
136	14	6	13	37	99
137	6	37	37	13	70
138	80	70	62	113	107
139	108	62	113	62	79
140	90	113	99	79	37
141	24	107	107	99	84
142	98	98	98	76	93
143	93	93	93	84	71
144	104	90	90	107	90
145	94	79	79	98	62
146	84	84	84	93	43
147	112	76	76	90	76
148	76	112	112	104	112
149	79	104	104	112	104

Tabla A.1: Tabla completa de características ponderadas por algoritmo

Apéndice B

Tabla completa de ponderación de características por normalización

Rank	**N° de carácterística**
1	126
2	3
3	138
4	140
5	127
6	125
7	124
8	149
9	121
10	4
11	139
12	137
13	147
14	102
15	30
16	32
17	46
18	128
19	44
20	60
21	100
22	88
23	2
24	86
25	58
26	91
27	143

28	135
29	122
30	123
31	141
32	134
33	35
34	21
35	77
36	117
37	18
38	116
39	74
40	114
41	87
42	59
43	49
44	72
45	31
46	101
47	136
48	132
49	45
50	105
51	148
52	63
53	16
54	131
55	115
56	120
57	7
58	109
59	73
60	17

61	81
62	129
63	11
64	25
65	39
66	106
67	118
68	67
69	64
70	78
71	53
72	10
73	119
74	95
75	29
76	145
77	38
78	47
79	50
80	92
81	22
82	1
83	66
84	52
85	8
86	146
87	108
88	80
89	61
90	142
91	36
92	12
93	69

94	33
95	20
96	23
97	133
98	24
99	82
100	130
101	111
102	27
103	19
104	40
105	9
106	110
107	57
108	103
109	65
110	26
111	97
112	56
113	94
114	54
115	68
116	144
117	5
118	55
119	96
120	75
121	34
122	83
123	43
124	28
125	42
126	89

127	14
128	15
129	85
130	48
131	41
132	71
133	113
134	13
135	51
136	6
137	99
138	70
139	37
140	107
141	62
142	98
143	93
144	79
145	90
146	84
147	76
148	104
149	112

Tabla B.1: Tabla completa de ponderación de características por normalización.