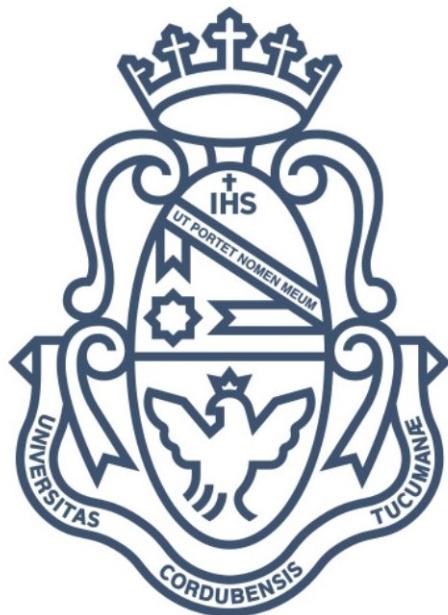


UNIVERSIDAD NACIONAL DE CÓRDOBA
FACULTAD DE CIENCIAS EXACTAS, FÍSICAS Y
NATURALES
FACULTAD DE CIENCIAS MÉDICAS
Ingeniería Biomédica

Proyecto Integrador
“Clasificación automática de densidad mamaria”

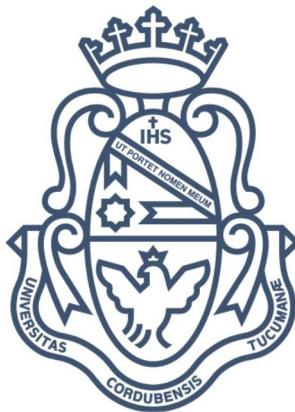


Bertero, Micaela Agustina
Griffa, Guillermmina

Córdoba, Diciembre, 2018

UNIVERSIDAD NACIONAL DE CÓRDOBA
FACULTAD DE CIENCIAS EXACTAS, FÍSICAS Y NATURALES
FACULTAD DE CIENCIAS MÉDICAS
Ingeniería Biomédica

Proyecto Integrador
“Clasificación automática de densidad mamaria”



Bertero, Micaela Agustina

Matrícula: 38280282

Griffa, Guillermmina

Matrícula: 38158691

Director: Valeria Rulloni

Asesores: Roberto Valfré y Elmer Fernández

Córdoba, Diciembre, 2018

Agradecimientos

En primer lugar nos gustaría agradecerle a nuestra Directora, Valeria Rulloni, por su paciencia y dedicación a lo largo del proyecto.

A nuestros asesores, Roberto Valfré y Elmer Fernandez, por su colaboración.

Al Instituto Oulton, por conferirnos la base de imágenes, haciendo una mención especial al Departamento de Bioingeniería.

A nuestras familias por su apoyo incondicional.

Resumen

La mamografía se considera en la actualidad una herramienta de diagnóstico muy útil para la detección de lesiones mamarias. Sin embargo, la sensibilidad de éstas disminuye conforme aumenta la densidad del tejido. Con el objetivo de determinar dicha densidad, en este trabajo se desarrollan distintos descriptores basados en análisis de brillo y textura, tales como histograma, entropía de permutaciones, dimensión fractal, etc. Los cuales constituirán luego el conjunto de entrada de los algoritmos de clasificación automática utilizados (SVC, k-NN, Random Forest, Regresión logística, entre otros).

El mejor resultado fue obtenido con el clasificador SVC, aunque todos se encuentran dentro del mismo rango. A su vez, mediante un análisis más específico, se llega a la conclusión de que un valor similar puede obtenerse reduciendo la cantidad de características a la entrada, lo que lleva a una reducción de los tiempos de cómputo.

Índice general

| | | |
|----------|--|----------|
| 1 | Introducción | 1 |
| 1.1 | Objetivo general | 2 |
| 1.2 | Objetivos específicos | 2 |
| 2 | Marco teórico | 3 |
| 2.1 | Anatomía de la mama | 3 |
| 2.2 | Tamizaje | 5 |
| 2.3 | Mamografía | 5 |
| 2.3.1 | Equipo | 7 |
| 2.3.2 | Sistema automático de control de exposición | 17 |
| 2.3.3 | Características generales de las imágenes mamográficas digitales | 18 |
| 2.3.4 | Herramientas de visualización | 18 |
| 2.4 | Densidad mamaria | 19 |
| 2.4.1 | Factores que influyen en la densidad mamaria | 21 |
| 2.4.2 | Clasificación de la densidad mamaria | 22 |
| 2.5 | Evaluación diagnóstica de la mama | 26 |
| 2.6 | Estándar DICOM | 26 |
| 2.7 | Procesamiento digital de imágenes | 27 |
| 2.7.1 | Relaciones básicas entre píxeles | 27 |
| 2.7.2 | Etapas fundamentales del procesamiento digital de imágenes | 28 |
| 2.8 | Preprocesamiento | 28 |
| 2.8.1 | Operaciones de punto | 29 |
| 2.8.2 | Operaciones morfológicas | 31 |
| 2.8.3 | Filtrado | 32 |
| 2.9 | Segmentación | 36 |
| 2.9.1 | Umbralización | 36 |
| 2.10 | Representación y descripción | 37 |
| 2.10.1 | Área | 38 |
| 2.10.2 | Operador gradiente | 38 |
| 2.10.3 | Histograma | 39 |

| | |
|---|-----------|
| 2.10.4 Entropía de permutaciones | 41 |
| 2.10.5 Análisis fractal | 42 |
| 2.10.6 Descriptores de textura de Haralick | 43 |
| 2.11 Reconocimiento de patrones | 46 |
| 2.11.1 Máquinas de vectores de soporte | 46 |
| 2.11.2 k-vecinos más próximos (k-NN) | 53 |
| 2.11.3 Regresión logística | 55 |
| 2.11.4 Random forest | 57 |
| 2.11.5 Análisis de componentes principales (ACP) | 57 |
| 2.12 Estado del arte | 60 |
| 2.12.1 Cuantificación de la densidad mamaria mediante Entropía de Permutaciones en mamografías [31] | 61 |
| 2.12.2 Breast Density Characterization using Texton Distributions[32] | 62 |
| 2.12.3 Automated analysis of mammographic densities [33] | 62 |
| 2.12.4 Análisis de la densidad de mama asistido por ordenador [34] | 62 |
| 2.12.5 Automated classification of parenchymal patterns in mammograms [35] | 63 |
| 2.12.6 Texture Descriptors applied to Digital Mammography [36] | 63 |
| 3 Materiales y métodos | 65 |
| 3.1 Materiales | 65 |
| 3.2 Etapas fundamentales | 66 |
| 3.3 Comienzos | 66 |
| 3.4 Segmentación | 68 |
| 3.4.1 Eliminación de etiqueta | 69 |
| 3.4.2 Sustracción de implante mamario | 70 |
| 3.4.3 Extracción del músculo pectoral | 71 |
| 3.4.4 Sustracción de piel | 77 |
| 3.5 Extracción de características | 79 |
| 3.5.1 Cuantificación de píxeles de la ROI | 79 |
| 3.5.2 Análisis de histograma global | 79 |
| 3.5.3 Análisis de histograma local | 81 |
| 3.5.4 Entropía de permutaciones | 86 |
| 3.5.5 Análisis fractal | 88 |
| 3.5.6 Descriptores de textura de Haralick | 91 |
| 3.5.7 Características obtenidas de la información DICOM | 91 |
| 3.6 Preprocesamiento de datos | 93 |
| 3.6.1 Análisis de características | 93 |

| | | |
|--|---|------------|
| 3.6.2 | Eliminación de datos faltantes | 100 |
| 3.6.3 | Codificación de variables categóricas | 101 |
| 3.6.4 | Escalamiento de los datos | 103 |
| 3.6.5 | Reducción de dimensionalidad | 104 |
| 3.7 | Ajuste de los estimadores | 104 |
| 3.7.1 | Validación cruzada | 104 |
| 3.7.2 | Selección de parámetros utilizando GridSearchCV | 105 |
| 4 | Resultados | 106 |
| 4.1 | Clasificación con la base de datos segmentada | 106 |
| 4.2 | Clasificación con la base de datos completa | 112 |
| 4.3 | Clasificación luego de aplicar ACP | 119 |
| 4.4 | Tratamiento del desbalance entre clases | 120 |
| 5 | Conclusiones | 122 |
| 6 | Trabajos futuros | 123 |
| Bibliografía | | 123 |
| A DDSM (The Digital Database for Screening Mammography) | | 127 |

Índice de tablas

| | |
|---|-----|
| Tabla 3.6.1 Descripción de las características del histograma local | 96 |
| Tabla 3.6.2 Descripción de las características extraídas de la información DICOM. | 99 |
| Tabla 3.6.3 Descripción de las características relacionadas con el operador gradiente. | 100 |
| Tabla 4.2.1 Comparación de los resultados obtenidos con los diferentes algoritmos. | 118 |
| Tabla 4.4.1 Comparación de los resultados obtenidos luego de triplicar los ejemplos de las clases “a” y “d” | 120 |

Índice de figuras

| | |
|--|----|
| Figura 2.1.1 Anatomía mamaria.[2] | 4 |
| Figura 2.3.1 Componentes del equipo de mamografía.[7] | 7 |
| Figura 2.3.2 Punto focal.[6] | 8 |
| Figura 2.3.3 Efecto talón.[6] | 9 |
| Figura 2.3.4 Si el cátodo es dirigido a la pared torácica, se puede aprovechar el efecto talón para producir una densidad óptica más uniforme.[6] | 10 |
| Figura 2.3.5 Al inclinar el tubo de rayos X en su carcasa, el punto focal disminuye y mejora la resolución espacial.[6] | 11 |
| Figura 2.3.6 Película radiográfica[6]. | 14 |
| Figura 2.3.7 Sistema de doble lectura de fósforo fotoestimulable[9] | 15 |
| Figura 2.3.8 Esquema del detector de doble capa de a-Se. PCL: Capa fotoconductiva; ETL: Capa de captura de electrones.[9] | 17 |
| Figura 2.4.1 Anatomía interna de la mama en una imagen radiográfica. A. Conducto galactóforo. B. lóbulos. C. Sección de un conducto galactóforo. D. Pezón. E. Tejido adiposo. F. Músculo pectoral mayor. G. Espacio Retromamario. H. Ligamentos de Cooper. I. Espacio retromamario. J. Piel. K. Folio inframamario.[8] | 21 |
| Figura 2.4.2 Mamografía digital con ACR a. | 24 |
| Figura 2.4.3 Mamografía digital con ACR b. | 24 |
| Figura 2.4.4 Mamografía digital con ACR c. | 25 |
| Figura 2.4.5 Mamografía digital con ACR d. | 25 |
| Figura 2.7.1 Definición de imagen. | 27 |
| Figura 2.7.2 Relaciones entre píxeles. [15] | 28 |
| Figura 2.7.3 Etapas fundamentales del procesamiento digital de imágenes | 28 |
| Figura 2.8.1 Binarización y complementación respectivamente. | 30 |
| Figura 2.8.2 Imagen negativa * 2000 + imagen original. | 31 |
| Figura 2.8.3 Imagen binaria, dilatada y erosionada. | 32 |
| Figura 2.8.4 Convolución. [17] | 33 |
| Figura 2.8.5 Kernels utilizados en técnica de realce de bordes por medio del gradiente direccional.[18] | 34 |
| Figura 2.8.6 Imagen resultante luego de aplicar filtro de gradiente. | 35 |

| | |
|--|----|
| Figura 2.8.7 Ejemplo de máscara para filtro gaussiano.[19] | 35 |
| Figura 2.9.1 Umbral único y umbral multinivel.[15] | 36 |
| Figura 2.10.1 $\nabla f(x_0, y_0) = (u, v)$ establece la dirección de máximo crecimiento de nivel de gris. | 38 |
| Figura 2.10.2 Representación del histograma de una imagen digital monocromática. | 39 |
| Figura 2.10.3 Asimetría positiva ($As > 0$). B) Simetría($As = 0$). C) Asimetría negativa ($As < 0$). Asimetría positiva ($As > 0$). B) Simetría($As = 0$). C) Asimetría negativa ($As < 0$). | 40 |
| Figura 2.10.4 Distribución leptocúrtica ($Cr > 0$). B) Distribución mesocúrtica ($Cr = 0$). C) Distribución platicúrtica ($Cr < 0$). | 40 |
| Figura 2.10.5 Celda de resolución.[25] | 43 |
| Figura 2.11.1 Hiperplanos de separación en un espacio bidimensional de un conjunto de ejemplos separables en dos clases: (a) ejemplo de hiperplano de separación (b) múltiples ejemplos de los infinitos posibles hiperplanos de separación. | 48 |
| Figura 2.11.2 Márgenes de hiperplanos de separación. a) hiperplano de separación no óptimo y su margen asociado (no máximo), b) hiperplano de separación óptimo y su margen asociado (máximo). | 48 |
| Figura 2.11.3 La distancia de cualquier ejemplo x_i al hiperplano de separación óptimo viene dada por $ D(x_i) /\ w\ $. En particular, si dicho ejemplo pertenece al conjunto de vectores soporte (identificados por siluetas sólidas), la distancia a dicho hiperplano será siempre $1/\ w\ $. Además, los vectores soporte aplicados a la función de decisión siempre cumplen que $ D(x) = 1$. | 49 |
| Figura 2.11.4 Obtención de hiperplano óptimo en ejemplos no separables. | 51 |
| Figura 2.11.5 El problema de la búsqueda de una función de decisión no lineal en el espacio de entradas, se puede transformar en un nuevo problema consistente en la búsqueda de una función de decisión lineal (hiperplano) en un nuevo espacio transformado (espacio de características). | 52 |
| Figura 2.11.6 Ejemplo de notación para algoritmo k-NN. | 54 |
| Figura 2.11.7 Pseudocódigo para el clasificador k-NN. | 54 |
| Figura 2.11.8 Ejemplo de aplicación del algoritmo k-NN básico. | 54 |
| Figura 2.11.9 Ejemplo de la no monotocidad del porcentaje de bien clasificados en función de K. | 55 |
| Figura 2.11.10 Ejemplo de la no monotocidad del porcentaje de bien clasificados en función de K. | 56 |
| Figura 2.11.11 Algoritmo Random Forest. [29] | 57 |
| Figura 2.11.12 Ejemplo de la recta que minimiza las distancias ortogonales de los puntos a ella. | 58 |
| Figura 2.12.1 Tejido glandular de 4 mamografías que representan las distintas densidades. Imágenes reales procesadas por los algoritmos.[31] | 61 |

| | |
|--|----|
| Figura 2.12.2Matriz de confusión de test.[31] | 61 |
| Figura 2.12.3Classification accuracy results. [32] | 62 |
| Figura 2.12.4Coeficientes de correlación intraclass, que resumen la performance de la clasificación automática versus la del radiólogo (R1). [33] | 62 |
| Figura 2.12.5Comparación métodos estudiados. [34] | 63 |
| Figura 2.12.6Las fracciones de error menores y mayores e1 y e2 de las clasificaciones se obtuvieron mediante diferentes combinaciones de características. [35] | 63 |
| Figura 2.12.7Porcentaje de clasificación correcta con y sin selección de características. [36] | 64 |
| | |
| Figura 3.3.1 Imagen mamográfica analógica digitalizada (izquierda) Vs. Imagen digital(derecha). | 67 |
| Figura 3.3.2 Imagen mamográfica analógica digitalizada luego de eliminar el ruido. | 67 |
| Figura 3.3.3 Extracción del músculo pectoral | 68 |
| Figura 3.4.1 Eliminación de etiqueta | 69 |
| Figura 3.4.2 Sustracción de implante mamario | 71 |
| Figura 3.4.3 Umbralización y ubicación de puntos generadores (“Semillas”). | 73 |
| Figura 3.4.4 Aplicación de operador gradiente y crecimiento de región. | 73 |
| Figura 3.4.5 Coordenadas de los puntos para el ajuste de curva. | 74 |
| Figura 3.4.6 Curva obtenida e imagen resultante. | 74 |
| Figura 3.4.7 Recta obtenida e imagen resultante. | 74 |
| Figura 3.4.8 Músculo se confunde con glándula. | 76 |
| Figura 3.4.9 Músculo poco intenso. | 76 |
| Figura 3.4.10Músculo bien definido. | 77 |
| Figura 3.4.11Extracción de músculo a imagen con implante mamario. | 77 |
| Figura 3.4.12Sustracción de piel. | 78 |
| Figura 3.5.1 Histograma de mamografía con densidad a. | 80 |
| Figura 3.5.2 Histograma de mamografía con densidad b. | 80 |
| Figura 3.5.3 Histograma de mamografía con densidad c. | 81 |
| Figura 3.5.4 Histograma de mamografía con densidad d. | 81 |
| Figura 3.5.5 División de la mama en regiones según la distancia a la piel. | 82 |
| Figura 3.5.6 Vectores obtenidos para mamografía de densidad a. | 82 |
| Figura 3.5.7 Vectores obtenidos para mamografía de densidad b. | 83 |
| Figura 3.5.8 Vectores obtenidos para mamografía de densidad c. | 83 |
| Figura 3.5.9 Vectores obtenidos para mamografía de densidad d. | 83 |
| Figura 3.5.10Relaciones obtenidas para mamografía de clase a. | 84 |
| Figura 3.5.11Relaciones obtenidas para mamografía de clase b. | 84 |
| Figura 3.5.12Relaciones obtenidas para mamografía de clase c. | 85 |
| Figura 3.5.13Relaciones obtenidas para mamografía de clase d. | 85 |

| | |
|---|-----|
| Figura 3.5.14 Entropías calculadas para mamografías de clase a | 87 |
| Figura 3.5.15 Entropías calculadas para mamografías de clase b | 87 |
| Figura 3.5.16 Entropías calculadas para mamografías de clase c | 87 |
| Figura 3.5.17 Entropías calculadas para mamografías de clase d | 87 |
| Figura 3.5.18 Imágenes segmentadas por medio del algoritmo de Otsu multiumbral. | 88 |
| Figura 3.5.19 Ajuste de recta en algoritmo de conteo de recuadros (Box Counting). | 89 |
| Figura 3.5.20 Vector con el cálculo de la dimensión fractal y de sus valores asociados. | 89 |
| Figura 3.5.21 Bordes de la imagen segmentada. | 90 |
| Figura 3.5.22 Aumento de los bordes de la imagen segmentada. | 90 |
| Figura 3.5.23 Ejemplo de la matriz de los descriptores de Haralick. | 91 |
| Figura 3.6.1 Variación del desvío estándar de los niveles de gris según la densidad mamaria. | 94 |
| Figura 3.6.2 Variación de la curtosis de los niveles de gris según la densidad mamaria. | 94 |
| Figura 3.6.3 Variación de la asimetría de los niveles de gris según la densidad mamaria. | 95 |
| Figura 3.6.4 Relación entre el tamaño de la ROI y la densidad radiológica. | 95 |
| Figura 3.6.5 Variación de la densidad mamaria con la edad. | 96 |
| Figura 3.6.6 Relación entre el kilo voltaje pico utilizado y la densidad mamaria. | 97 |
| Figura 3.6.7 Relación entre la corriente (mA) del tubo de rayos X y la densidad mamaria. | 97 |
| Figura 3.6.8 Relación entre la fuerza con la que se comprime la mama durante el estudio y la densidad mamaria. | 98 |
| Figura 3.6.9 Relación entre el espesor de la mama comprimida durante el estudio y la densidad mamaria. | 98 |
| Figura 3.6.10 Relación entre el promedio del módulo del gradiente y la densidad mamaria. | 99 |
| Figura 3.6.11 Diferencia en la relación de (pixels blancos/total de pixels) de la imagen con el filtro de gradiente aplicado con la densidad mamaria. | 100 |
| Figura 3.6.12 Atributos que contienen la mayor cantidad de datos faltantes. | 101 |
| Figura 3.6.13 Variables categóricas. | 102 |
| Figura 3.6.14 Ejemplos de variables categóricas codificadas. | 103 |
| Figura 3.7.1 Validación cruzada. | 105 |
| Figura 4.1.1 Reporte de la clasificación utilizando características extraídas del histograma y k-NN. | 107 |
| Figura 4.1.2 Matriz de confusión. | 107 |
| Figura 4.1.3 Reporte de la clasificación utilizando características extraídas de la información DICOM y k-NN. | 108 |
| Figura 4.1.4 Matriz de confusión. | 108 |
| Figura 4.1.5 Reporte de la clasificación utilizando características extraídas de la imagen a la que se le aplicó el operador gradiente y k-NN. | 108 |

| | |
|--|-----|
| Figura 4.1.6 Matriz de confusión. | 109 |
| Figura 4.1.7 Reporte de la clasificación utilizando los descriptores de textura de Haralick de la imagen y k-NN. | 109 |
| Figura 4.1.8 Matriz de confusión. | 110 |
| Figura 4.1.9 Reporte de la clasificación utilizando las características relacionadas con la entropía de permutaciones y k-NN. | 110 |
| Figura 4.1.10Matriz de confusión. | 110 |
| Figura 4.1.11Reporte de la clasificación utilizando las características relacionadas con la dimensión fractal de la imagen y k-NN. | 111 |
| Figura 4.1.12Matriz de confusión. | 111 |
| Figura 4.1.13Reporte de la clasificación utilizando las características relacionadas con la dimensión fractal de los bordes de la de la imagen y k-NN. | 111 |
| Figura 4.1.14Matriz de confusión. | 112 |
| Figura 4.2.1 Reporte de la clasificación utilizando Random Forest. | 113 |
| Figura 4.2.2 Matriz de confusión para la clasificación con Random Forest. | 113 |
| Figura 4.2.3 Matriz de confusión de los errores para la clasificación con Random Forest. . | 114 |
| Figura 4.2.4 Reporte de la clasificación utilizando Regresión Logística. | 114 |
| Figura 4.2.5 Matriz de confusión para la clasificación con Regresión Logística. | 115 |
| Figura 4.2.6 Matriz de confusión de los errores para la clasificación con Regresión Logística. | 115 |
| Figura 4.2.7 Reporte de la clasificación utilizando k-NN. | 116 |
| Figura 4.2.8 Matriz de confusión para la clasificación con k-NN. | 116 |
| Figura 4.2.9 Matriz de confusión de los errores para la clasificación con k-NN. | 117 |
| Figura 4.2.10Reporte de la clasificación utilizando SVC. | 117 |
| Figura 4.2.11Matriz de confusión para la clasificación con SVC. | 118 |
| Figura 4.2.12Matriz de confusión de los errores para la clasificación con SVC. | 118 |
| Figura 4.3.1 Experimentación con seis diferentes algoritmos. | 119 |
| Figura 4.3.2 Reporte de la clasificación utilizando ACP y SVC. | 119 |
| Figura 4.3.3 Matriz de confusión para la clasificación con ACP y SVC. | 120 |
| Figura 4.4.1 Reporte de la clasificación con SVC. | 121 |
| Figura 4.4.2 Matriz de confusión para la clasificación con SVC. | 121 |

Capítulo 1

Introducción

Las pautas actuales del Departamento de Salud y Servicios Humanos de los Estados Unidos (HHS) y el Colegio Estadounidense de Radiología (ACR) recomiendan realizarse una mamografía de exploración cada año en las mujeres, comenzando a partir de los 40 años. Este estudio juega un papel central en la detección temprana del cáncer de mama.

La densidad radiográfica de la mama se refiere a la cantidad de tejido parenquimatoso y conectivo que, por ser radiolúcido, aparece de color blanco. El tejido canceroso posee el mismo comportamiento al interaccionar con la radiación, es por ello que los tumores pueden ser difíciles de percibir entre los tejidos densos. Por esta razón, la sensibilidad de la mamografía disminuye a medida que la densidad aumenta. A su vez, las mujeres con tejido mamario denso parecen tener un riesgo ligeramente mayor de padecer cáncer de seno.

Los métodos iniciales para evaluar la densidad mamográfica fueron completamente subjetivos y cualitativos; sin embargo, en los últimos años se han desarrollado métodos para proporcionar mediciones más objetivas, y a eso nos abocaremos a lo largo de este proyecto.

Los radiólogos utilizan el sistema BI-RADS para clasificar la densidad mamaria en cuatro categorías que van desde tejido adiposo casi en su totalidad, hasta tejido extremadamente denso con muy poca grasa. Basándonos en esta escala, hemos decidido construir un algoritmo que sea capaz de clasificar la densidad a partir de una imagen de mamografía digital.

La realización del proyecto puede ser dividido en cuatro etapas:

- Construcción de una base de datos a partir de imágenes digitales, que han sido previamente clasificadas por especialistas, extraídas del Instituto Oulton.
- Construcción de algoritmo de segmentación de las imágenes para obtener la región de interés (eliminación de zonas irrelevantes en el análisis: músculo pectoral, piel y prótesis).
- Extracción de características de la región de interés mediante técnicas de procesamiento de imágenes basadas en niveles de brillo y análisis de textura.
- Utilización de algoritmos inteligentes de agrupamiento.

1.1. Objetivo general

Este proyecto integrador tiene como objetivo general clasificar mamografías digitales según su densidad radiológica haciendo uso de algoritmos inteligentes de agrupamiento.

1.2. Objetivos específicos

1. Obtener características de interés en mamografías digitales mediante el uso de técnicas de procesamiento de imágenes.
2. Comprobar que existe una relación entre la textura de la imagen y su densidad radiológica.
3. Armado de una base de datos adecuada para el entrenamiento de la metodología automática.
4. Comprobar si la cantidad de coincidencias entre la evaluación del médico y nuestro sistema aumenta al incrementar la cantidad de características extraídas de la región de interés.

Capítulo 2

Marco teórico

2.1. Anatomía de la mama

La mama es una glándula sudorípara apocrina modificada que se desarrolla en la pubertad. Se ubica en la pared torácica anterior, recubriendo el músculo pectoral mayor entre la segunda y la sexta costilla verticalmente y desde el esternón medialmente hasta la línea axilar media lateralmente. El tejido mamario se extiende hacia la axila baja como una proyección de forma triangular: esta parte de la mama se llama cola axilar o Cola de Spence.

Es posible describir la estructura de la mama dividiéndola principalmente en cuatro porciones, de exterior a interior:[1]

- *Revestimiento cutáneo:*

La mama está cubierta por una piel fina y móvil que continúa en la periferia con la piel del tórax. Su vértice está constituido por un área redondeada y pigmentada denominada areola, en cuya superficie se localizan unas glándulas sebáceas denominadas glándulas areolares y en el centro se encuentra el pezón.

- *Tejido subcutáneo:*

Una capa adiposa de tejido subcutáneo se extiende por la cara interior de la piel de la mama, excepto a nivel de la areola y del pezón. El tejido subcutáneo está tabicado por hojas conjuntivas fibrosas que se extienden desde la cara interior de la dermis, hasta la cara anterior de la glándula mamaria y los conductos galactóforos, sobre la cual se insertan. En ciertas regiones este tejido se condensa para formar los ligamentos suspensorios de la mama o ligamentos de Cooper. Estos ligamentos limitan, entre la piel y la glándula celdas ocupadas por tejido adiposo: fosas adiposas.

- *Tejido glandular:*

La glándula mamaria está constituida por 15 a 20 lóbulos que irradian desde el pezón, cada uno de los cuales tiene independencia funcional. Éstos, a su vez, están formados por la unión

de numerosos lobulillos donde se encuentran los acinos, encargados de la producción de leche durante el embarazo y lactancia. Cada uno de los lóbulos posee un conducto excretor: el conducto galactóforo, que antes de llegar al pezón presentan una dilatación: el seno galactóforo. A la unión del lóbulo con su respectivo conducto galactóforo se lo denomina unidad ductolobular terminal.

■ *Capa adiposa retromamaria:*

En la cara posterior de la glándula existe una capa de tejido adiposo. Detrás de la capa retromamaria se encuentra la capa membranosa del tejido subcutáneo, que está separada de la fascia del músculo pectoral mayor por tejido adiposo laxo, esto permite un ligero movimiento del pecho sobre la pared torácica.

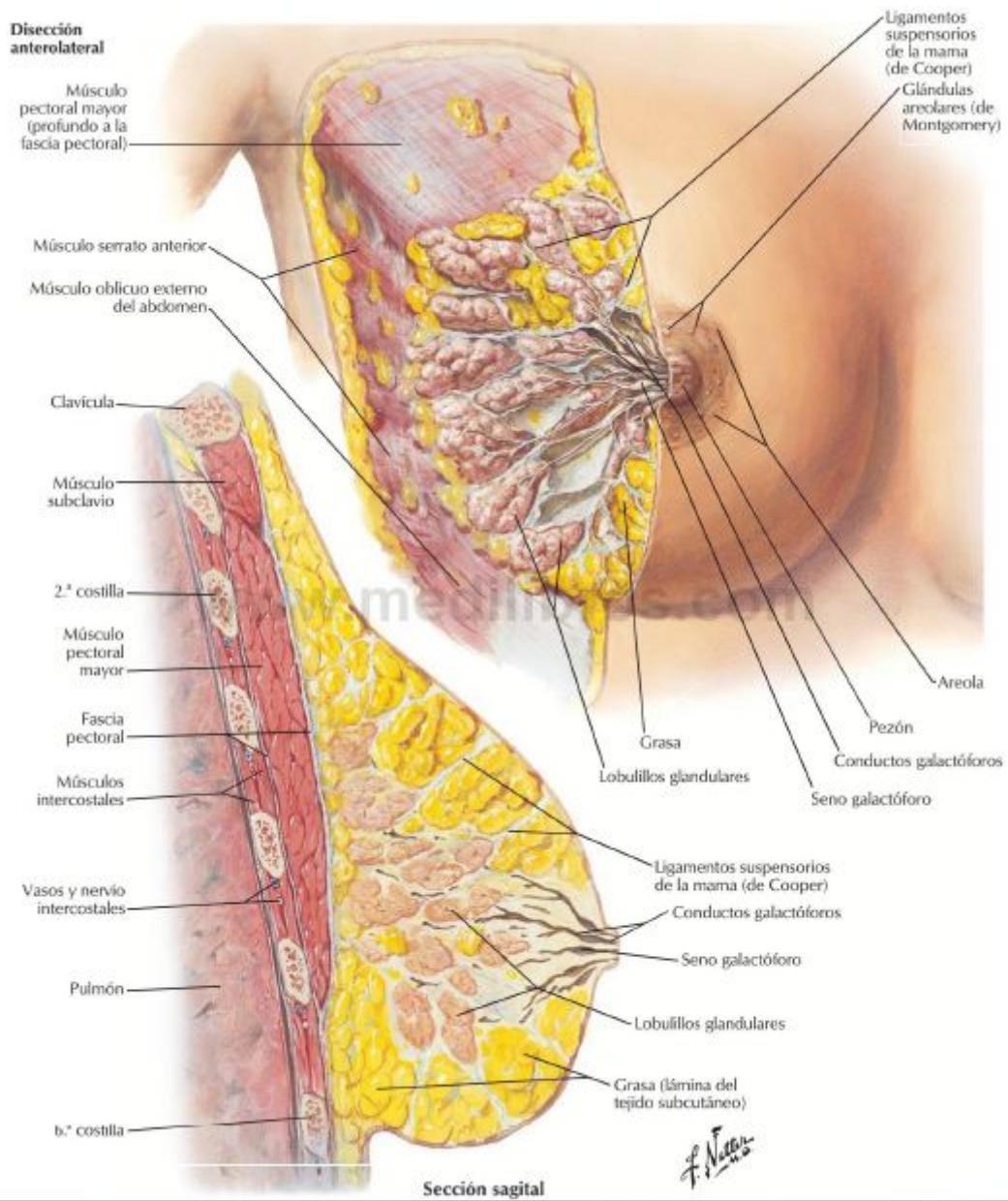


Figura 2.1.1: Anatomía mamaria.[2]

2.2. Tamizaje

El cáncer de mama es la primera causa de muerte por cáncer en mujeres argentinas: se producen más de 5.800 muertes por año por esta enfermedad. Es el cáncer de mayor incidencia en mujeres: representa el 31.8 % de los mismos.¹

Más del 75 % de las mujeres con cáncer de mama no tienen ningún antecedente familiar de dicha enfermedad [3]. Por lo cual, la detección temprana a fin de mejorar el pronóstico y la supervivencia sigue siendo la piedra angular de la lucha contra este cáncer, para ello existen dos métodos:

- El diagnóstico precoz o el conocimiento de los primeros signos y síntomas en la población sintomática, para facilitar el diagnóstico y el tratamiento temprano.
- El cribado, es decir, la aplicación sistemática de pruebas de tamizaje en una población aparentemente asintomática. Su objetivo es detectar a las personas que presenten anomalías indicativas de cáncer.

Los estudios clínicos aleatorizados sobre el tamizaje mamográfico han comprobado su eficacia, dado que han demostrado una disminución significativa de la mortalidad por cáncer de mama. Según datos de la Organización Mundial de la Salud (OMS), esta forma de cribado puede reducir la mortalidad por cáncer de mama en un 20 %-30 % en las mujeres de más de 50 años, si su cobertura supera el 70 %. [4]

Si bien la mamografía detecta la mayor parte de los tumores mamarios malignos, puede ocurrir que no se alcancen a percibir en las imágenes pero sí que sean palpables. Por consiguiente, el Colegio Americano de Radiología (American College of Radiology, ACR) considera que la palpación mamaria constituye una parte importante del tamizaje.

Por definición, el tamizaje mamográfico consiste en tomar una proyección oblicua mediolateral y una craneocaudal de cada mama con el objetivo de localizar al pequeño subgrupo de mujeres que deben realizarse más estudios por imágenes, del amplio grupo de mujeres a quienes se recomienda el tamizaje periódico. En algunos entornos clínicos, se toman más proyecciones mamográficas o se realizan otros estudios por imágenes de la mama en la misma cita, si surge alguna duda en las imágenes de tamizaje. No obstante, es más frecuente que la interpretación de las imágenes de tamizaje se realice por grupos, con lo cual se vuelve a citar a las mujeres que presentan alguna imagen anómala para efectuarles nuevos estudios en un momento posterior.[5]

2.3. Mamografía

Las tecnologías actualmente disponibles para la obtención de imágenes mamarias se utilizan para identificar diferencias estructurales o morfológicas en tumores, como microcalcificaciones, masas ti-

¹Ministerio de Salud de la Nación. <http://www.msal.gov.ar/inc/acerca-del-cancer/cancer-de-mama/>.

sulares, angiogénesis, asimetría y distorsión arquitectónica. Algunas de las técnicas desarrolladas más recientemente pueden proporcionar información sobre las diferencias en las características biológicas o funcionales entre tumores y tejidos normales. Sin embargo, hasta ahora no hay una única modalidad que pueda lograr simultáneamente todos estos objetivos.

La mamografía se basa en la atenuación diferencial de los fotones de rayos X de los tejidos, sin embargo, para este caso, las técnicas radiográficas usuales son completamente inútiles y se diseñan técnicas específicas destinadas a incrementar la absorción diferencial. Esto es así debido a que se necesita alta resolución de contraste, por tratarse de tejidos con densidades y composiciones similares, de muy baja absorción, y con un contraste intrínseco muy pequeño. Además, se busca obtener una imagen con alta resolución espacial para poder detectar microcalcificaciones.

Para ello se requiere del uso de radiaciones de baja energía para las cuales aumenta la absorción fotoeléctrica, en el rango de 20 a 30 keV. Para energías superiores a 40 keV, los coeficientes lineales de atenuación de rayos X de los tejidos graso y fibroglandular se superponen, el carcinoma y el tejido glandular normal no se pueden diferenciar y el diagnóstico se ve comprometido debido al contraste deficiente de imagen.^[6]

Sin embargo, la reducción de la tensión de pico está ligada directamente a la capacidad de penetración del haz, lo que implica que se debe incrementar la corriente instantánea y con ella la dosis recibida ya que la mayoría de los fotones del haz son absorbidos en la mama. Como solución de compromiso se opta por un valor entre de 24 y 28 keV, no tan bajo como para inducir una dosis de radiación excesiva y no tan elevado como para deteriorar la calidad de la imagen.

Las sutiles propiedades de atenuación de rayos X entre los tejidos normales y cancerosos y los riesgos asociados a la radiación ionizante exigen técnicas de imágenes que minimicen la dosis y optimicen la calidad de la imagen. Esto promueve el refinamiento de equipos de rayos X dedicados para mamografías.

El contraste se define como la diferencia de densidades ópticas entre un objeto y su fondo. En la mamografía dependemos principalmente del contraste para distinguir el parénquima normal de las masas y distorsiones estructurales. Los factores que influyen de manera importante en el contraste de la imagen son:

- La calidad del haz utilizado (material del ánodo, kVp, tipo de generador y filtros).
- La fracción de radiación dispersa que alcanza la película (rejilla, tamaño de la masa y compresión) o detector (mamografía digital).
- El contraste de la película (conjunto pantalla -papel, procesado de la película) y densidad óptica del fondo.

La visualización de microcalcificaciones o estructuras inferiores a las décimas de milímetros está relacionada con la borrosidad o resolución de la imagen. Y depende principalmente de la geometría

utilizada (tamaño del foco y magnificación) del receptor de la imagen (conjunto pantalla-película y la grilla), de la borrosidad cinética (tiempo de disparo y compresión) y del contraste de la imagen.

El movimiento de la mama produce una falta de definición en la imagen denominada borrosidad cinética que se ve favorecido por un tiempo de exposición prolongado y una compresión deficiente.

2.3.1. Equipo

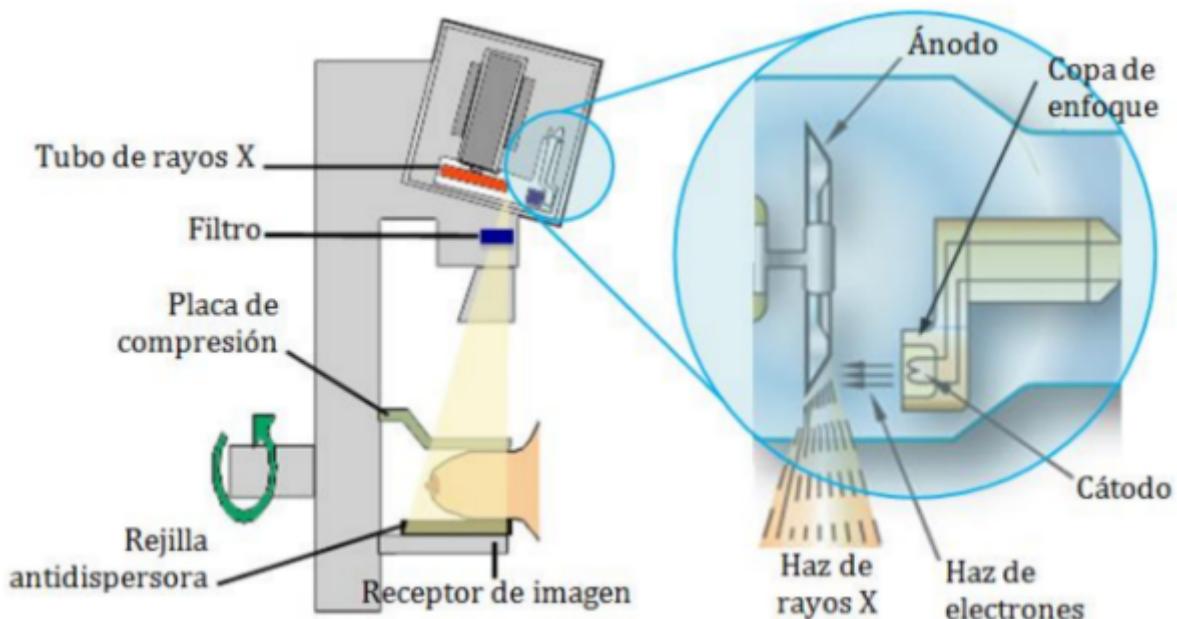


Figura 2.3.1: Componentes del equipo de mamografía.[\[7\]](#)

La mamografía se realiza con un equipo dedicado, generalmente con un brazo en forma de C destinado a facilitar el posicionamiento de los senos. El brazo C se puede ajustar en altura y orientación angular para ajustar la pala de compresión y el soporte del pecho a la posición de paciente sentado o parado. El tubo de rayos X y el conjunto de la mesa del receptor digital se montan en oposición: el tubo de rayos X para la generación del haz de fotones en la cabeza superior, un protector facial, una pala de compresión y el sistema de receptor de imagen en la parte inferior del brazo.

El sistema de adquisición de imágenes se compone de un tubo de rayos X, una placa de compresión mamaria y un sistema receptor de imágenes. La distancia desde el foco de rayos X hasta la plataforma de soporte de senos suele ser de alrededor de 60 cm. Normalmente se utiliza una rejilla anti-dispersión móvil que está situada justo detrás de la parte superior de la mesa de baja attenuación (a menudo de fibra de carbono) y frente al receptor de imagen.

Blanco

La principal diferencia en el tubo de rayos X es la composición del blanco, ya que para esta aplicación se utilizan materiales como el wolframio, el molibdeno o el rodio, especialmente estos

últimos dos. El motivo surge en función del espectro de emisión, tanto del pico característico como el de frenado.

Punto focal

El tamaño del punto focal también es una característica de máxima importancia debido a los requisitos de resolución muy alta, los tamaños de puntos focales de rayos X deben ser pequeños. Se utilizan puntos focales de aproximadamente $0,3 \times 0,3$ mm para la mamografía convencional, con un tamaño de $0,15 \times 0,15$ mm seleccionable para vistas ampliadas, donde la mama se aleja del receptor de imagen en una tabla de aumento especial para producir un aumento geométrico.

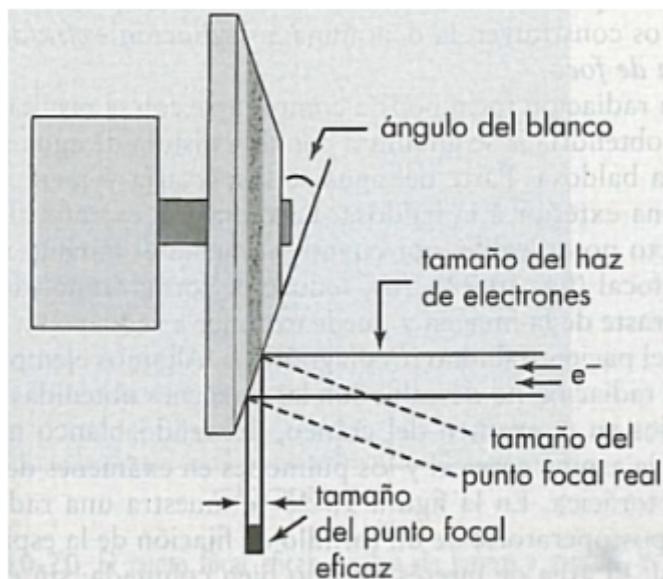


Figura 2.3.2: Punto focal.[6]

Efecto talón

El efecto talón es la variación de la intensidad de la radiación emitida, dependiendo del ángulo con que se emite respecto al ánodo. La intensidad del haz disminuye rápidamente desde el rayo central hasta el ánodo, debido en parte a que los rayos producidos a una pequeña profundidad del ánodo deben atravesar un mayor espesor hasta la superficie y por ello se atenúan.

En consecuencia, en las imágenes radiográficas, la parte del objeto situada en el lado anódico puede aparecer una mayor capacidad de atenuación, al ser de menor energía la radiación que incide en esta zona.

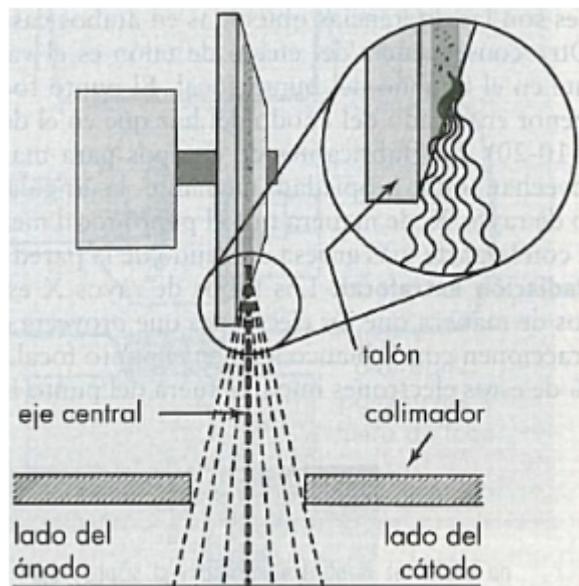


Figura 2.3.3: Efecto talón.[\[6\]](#)

La diferencia en la intensidad de radiación a través del haz útil de un campo de rayos X puede variar hasta en un 45 %. El rayo central del haz útil es la línea imaginaria generada por el rayo más centrado del haz. Si la intensidad de radiación a lo largo del rayo central se designa como el 100 %, entonces la intensidad en el cátodo puede ser hasta del 120 %, y sobre el ánodo puede ser tan baja como del 75 %.

El efecto talón origina una reducción sobre el ánodo de la intensidad de rayos X del haz útil debido a la absorción en el “talón” del blanco.

El efecto talón es importante cuando se muestran estructuras anatómicas que difieren mucho en espesor o en masa, como es el caso de la mamografía debido a la forma cónica de la mama, si bien hoy en día con el mecanismo de compresión se garantiza la uniformidad del grosor, en general, el posicionamiento del cátodo del tubo de rayos X sobre la parte más gruesa de la anatomía proporciona una exposición a la radiación del receptor de imagen más uniforme.

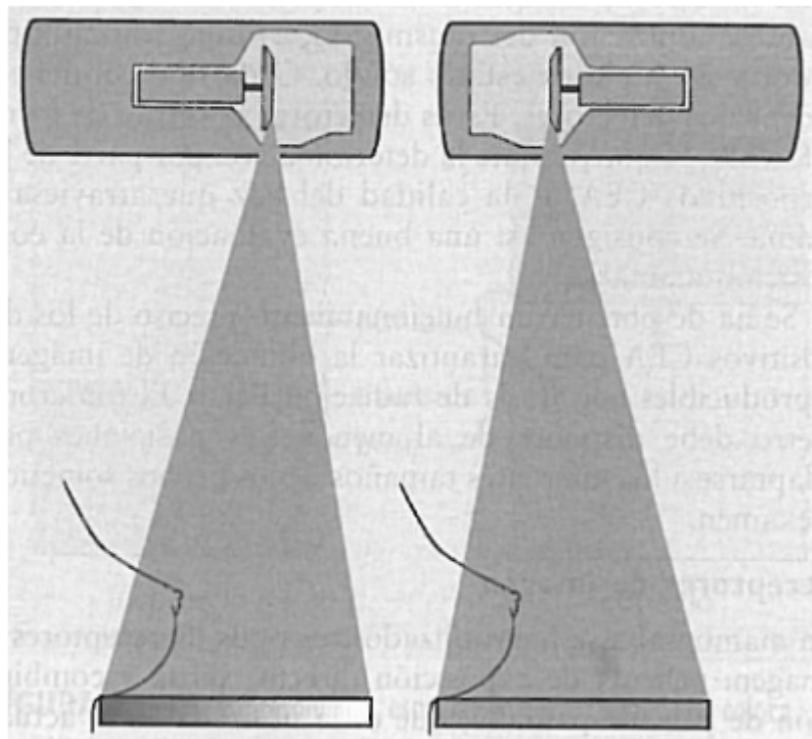


Figura 2.3.4: Si el cátodo es dirigido a la pared torácica, se puede aprovechar el efecto talón para producir una densidad óptica más uniforme.[\[6\]](#)

Al angular el tubo de rayos X, se aprovecha la ventaja del acortamiento que se produce en tamaño del punto focal, dando lugar a un punto focal efectivo aún menor. Otra consecuencia importante del efecto talón es que el tamaño del punto focal cambia. El punto focal efectivo es más pequeño en la parte del ánodo del tubo de rayos X que en la parte del cátodo. Algunos fabricantes de equipos de mamografías obtienen ventajas de esta propiedad inclinando el tubo de rayos X para producir puntos focales más pequeños a lo largo de la pared del tórax.

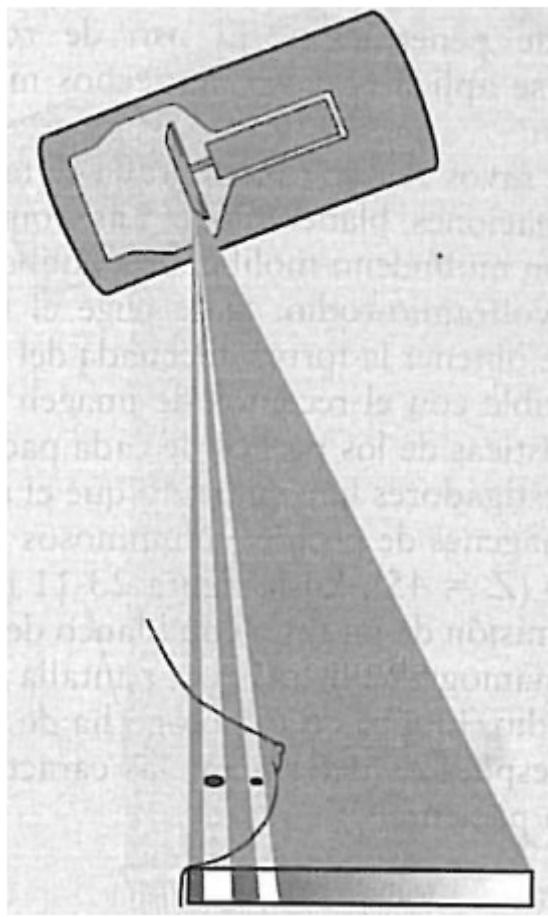


Figura 2.3.5: Al inclinar el tubo de rayos X en su carcasa, el punto focal disminuye y mejora la resolución espacial.[\[6\]](#)

El tubo de rayos X se coloca dentro de la unidad de modo que se emplee el efecto de talón del ánodo para reducir la intensidad de los rayos X hacia el lado del pezón del campo donde la mama será más delgada. Se confía mucho en el sistema de exposición automático de las unidades de mamografía modernas.

Espectro mamográfico de rayos X

El espectro de rayos X de un tubo con filtro de aluminio, encapsulado con vidrio y blanco de tungsteno convencional no es óptimo para la mamografía. Se considera que el mejor contraste entre el tejido normal y el tejido maligno se produce a una energía fotónica de alrededor de 20 keV. El aumento de la energía fotónica reducirá el contraste y la reducción de la energía fotónica dará lugar a una penetración inadecuada de la mama y un gran aumento en la dosis para el paciente, por lo que el espectro de rayos X es crítico. Una gama de espectros mamográficos son utilizados para mamografía digital. El objetivo del tubo de rayos X puede ser fácilmente commutable (dependiendo del diseño) entre el molibdeno y el rodio, o el rodio y el tungsteno, y el tubo tiene una ventana de salida de berilio de baja atenuación. Luego, el haz se filtra de molibdeno, rodio, plata o aluminio. El tubo de rayos X

funciona a una tensión en el rango de 25–35 keV.

Para senos más grandes, un rayo más penetrante es óptimo para evitar tiempos de exposición muy largos que brindan la posibilidad de borrosidad por movimiento, sobrecarga del tubo y altas dosis de radiación.[8]

Compresión mamográfica

En la mamografía, la mama se comprime utilizando una pala de compresión de plástico rígido transparente que puede ser impulsada por un motor. El uso de la fuerza de compresión reduce el grosor del seno y lo mantiene en su lugar, lo cual trae varias ventajas:

- Mejor resolución espacial. El seno se acerca al receptor de imágenes para reducir la ampliación y el desenfoque de las manchas focales.
- Movimiento borroso reducido, incluso en los tiempos de exposición relativamente largos (1s típico) comunes en la mamografía.
- Menos radiación dispersa en la imagen. La longitud de la trayectoria del haz a través del seno es más corta, por lo que hay menos material para hacer la dispersión. Reducir la proporción de radiación dispersada en la imagen mejora el contraste de la imagen y reduce el ruido de la imagen.
- Mejora la uniformidad de la imagen. La compresión extiende el tejido mamario de manera más uniforme a través de la imagen y hace que la patología sea más fácil de detectar.
- La reducción del grosor del seno comprimido disminuye el tiempo de exposición y disminuye la dosis de radiación administrada al seno.
- La longitud reducida de la trayectoria hace posible el uso de espectros de rayos X de menor energía (menos penetrante). Esto da mayor contraste.
- Las pequeñas áreas de patología enterradas en el tejido glandular se pueden visualizar mejor, ya que los tejidos malignos tienden a ser más firmes.

Las unidades de mamografía modernas pueden emplear un sistema para medir la cantidad creciente de fuerza que resulta de un pequeño aumento en la compresión para detener el movimiento motorizado en una compresión determinada. Muchas unidades utilizan un ensamblaje accionado por motor con más o menos compresión aplicada por el profesional; el técnico tiene control directo sobre la cantidad de compresión aplicada. El límite máximo de fuerza de compresión establecido en los sistemas de mamografía es de 200 Newtons.[8]

Detectores de imagen

Es posible clasificar a los detectores en tres categorías, según la tecnología del equipo mamográfico en cuestión.

Combinación película-pantalla

Para la mamografía analógica se han diseñado receptores de imágenes específicos, conformados por un conjunto de pantallas intensificadoras y películas especiales. La combinación de ambas se compone de películas de una sola emulsión adaptadas a una única pantalla.

Independientemente del tipo de película-pantalla que se utilice, ambas deben corresponderse espectralmente y suelen usarse emulsiones especiales acopladas a tierras raras. Los rayos X interactúan con la superficie de entrada de la pantalla, por lo tanto se coloca la película entre el tubo y la pantalla, con la emulsión cerca de esta última se mejora la resolución espacial. Cuando el haz de rayos X atraviesa a la paciente, la atenuación de los distintos tejidos produce variaciones en la radiación transmitida. Tras el paciente se forma un relieve de intensidad de rayos X conocido como imagen radiológica primaria. Como el ojo es insensible a los rayos X, esta imagen hay que convertirla en otra visible mediante pantallas fluorescentes, intensificadores de imagen, películas, etc.

Al alcanzar los rayos X la película radiográfica, la mayor parte la atraviesan y solamente una pequeña fracción, del orden del 1 % es absorbida por ella. Esto implica, que para formar la imagen es necesaria una dosis de radiación elevada. Para evitar este inconveniente se utilizan pantallas reforzadoras cuyo modo de funcionamiento es el siguiente: los fotones de rayos X son capturados por la pantalla. Luego, la energía absorbida en este proceso se emite en forma de fotones de luz visible (azul y verde), mediante un fenómeno de fluorescencia, que presenta una ganancia considerable en la relación del número de fotones emitidos (luz) al de absorbidos (rayos X), y por último, los fotones luminosos se transmiten a la película donde son absorbidos formando la imagen.

Esta combinación se coloca en un casete, chasis o portafilm, provisto de una sola pantalla, cuya capa de fósforo queda orientada hacia el tubo de rayos X. La película con emulsión en una sola cara, se sitúa con esta cara hacia la pantalla y su dorso, (parte no emulsionada) hacia el tubo. Dado que el haz utilizado en mamografía es de muy poca penetración el material utilizado para la construcción de los portafilm en mamografía debe ser de número atómico muy bajo, el más utilizado es fibra de carbono ya que deben cumplir dos requisitos indispensables ser sólidos y de material de escasa absorción. Es fundamental que el espesor de los mismos sea uniforme, no sólo entre sí, sino unos con otros para evitar las diferencias de exposición en mamas de similar composición y espesor originando zonas sobre o subexpuestas.

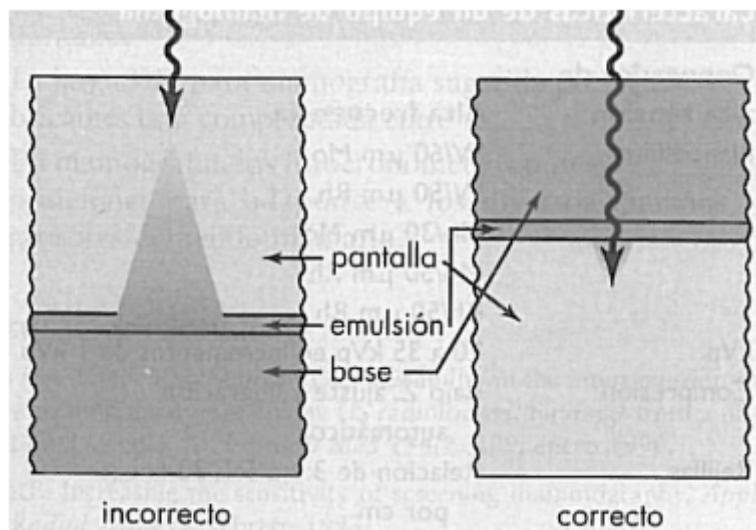


Figura 2.3.6: Película radiográfica[6].

El conjunto pantalla película se coloca en su interior por lo que presenta un diseño de libro, con una parte frontal de baja absorción a la radiación y una parte posterior opaca formando un sistema hermético a la luz. Su diseño debe asegurar:

- Un contacto perfecto entre pantalla y película.
- Una protección perfecta de la película siendo estanco a la luz.
- Un sistema de identificación de paciente.

Detectores de fósforo fotoestimulable

Los sistemas de CR consisten en placas de fósforos fotoestimulables que se introducen dentro de un chasis similar al del sistema película pantalla. La energía de los fotones de rayos X incidentes sobre la placa es absorbida localmente por los electrones de la red cristalina que pasan a niveles de energía superior metaestables donde quedan atrapados (centros F) formando la imagen latente, estable durante varias horas. Durante el proceso de lectura de la placa, un haz de luz láser muy focalizado realiza un barrido estimulando a los electrones a retornar al nivel de energía más bajo o nivel fundamental (luminiscencia estimulada). El paso a este nivel es realizado a través de transiciones entre niveles energéticos intermedios asociados a un material dopante que es introducido en la red cristalina. La longitud de onda de los fotones de luz emitidos depende del dopante utilizado y es distinta de la asociada a la luz láser. El número de fotones de luz emitidos en este proceso es proporcional al número de fotones de rayos X incidentes sobre la placa. La luz emitida es recogida por un fotomultiplicador donde se produce la conversión en señal eléctrica y su ulterior amplificación y digitalización. Para mejorar la eficiencia de recolección de la luz se ha desarrollado un sistema de doble lectura consistente en recoger la luz emitida por los fósforos por ambos lados de la placa. En los sistemas de CR, una

vez adquirida la imagen, es preciso borrar la información residual, lo que se consigue normalmente mediante un barrido de todo el fósforo con un haz de luz intensa que vacíe las trampas electrónicas.[\[9\]](#)

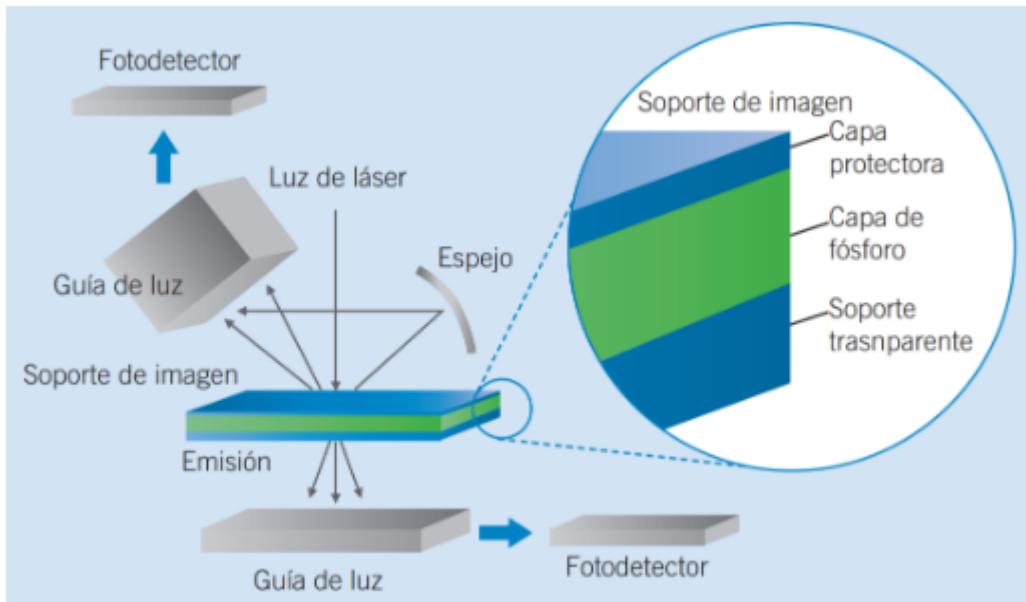


Figura 2.3.7: Sistema de doble lectura de fósforo fotoestimulable[\[9\]](#)

Detectores integrados

Estos sistemas, conocidos como sistemas rigurosamente digitales, tienen integrados el equipo de rayos X y el detector. Existen diferentes alternativas, pero al margen de las diferencias tecnológicas entre las mismas, todos ellos presentan ventajas notables entre las que cabe citar:

- Producen una imagen inmediata, sin procesos intermedios de revelado, de lectura ni de ningún otro tipo. Hacén desaparecer los chasis y, con ello, permiten construir un entorno puramente digital, limpio, con tiempos muertos menores y con capacidad para incrementar el rendimiento de salas y equipos.
- Reducen la dosis a los pacientes o al menos no la incrementan dado que la eficiencia de los detectores empleados es sensiblemente mayor.
- Producen imágenes de calidad muy apreciable, mucho más estable y con posibilidades muy grandes de adaptación a cada necesidad concreta. En particular, su resolución de contraste es muy superior a la de los sistemas convencionales.
- Es posible optimizar la cadena completa de obtención de imágenes incluyendo los factores de exposición (kVp, combinación ánodo-filtro, etc.) que se seleccionan en función de las características de la mama (atenuación y espesor). Una consecuencia de ello ha sido el ahorro importante en las dosis en pacientes con mamas gruesas al seleccionar de forma automática espectros de mayor energía.

Detectores de panel plano de selenio amorfo (a-Se)

Son los más utilizados en la actualidad para equipos de mamografía digital. El material utilizado habitualmente en la fabricación de este tipo de detectores es un fotoconductor que convierte directamente los fotones de rayos X en pares electrón-hueco. Por este motivo son denominados también de “conversión directa”. La carga generada es almacenada y medida como una señal electrónica. Los detectores de este tipo más extendidos son los fabricados con selenio amorfo (a-Se) como material fotoconductor. El haz de rayos X transmitido por la mama es absorbido por la capa de selenio generando pares electrón-hueco. La carga generada es recogida aplicando un campo eléctrico intenso entre un par de electrodos situados en las superficies superior e inferior de la capa de selenio.

Este método minimiza la dispersión lateral de la carga, garantizando una imagen de gran nitidez, con una función de transferencia de modulación alta, a la vez que se mantiene un espesor de detector suficiente para garantizar una eficiencia de detección también alta. La carga es leída por una matriz activa de TFT en contacto directo con la superficie inferior del selenio. En estos detectores es importante evitar la inyección de cargas desde los electrodos. Con este fin, se interpone una capa de bloqueo entre los electrodos y el material activo. El espesor de la capa de bloqueo se fija para impedir que queden atrapadas cargas entre dicha capa y el electrodo de lectura que, de estar presentes, dan lugar a una imagen latente que degrada la imagen final producida por el detector.

Este diseño reduce prácticamente a cero la corriente oscura del detector. Suelen indicarse como ventajas relativas de los sistemas de selenio amorfo su mayor eficiencia cuántica en la detección de los rayos X en el intervalo de las energías típicas utilizadas en mamografía (20 keV-30 keV).

La limitación clásica que se atribuye a los detectores de selenio es una cierta remanencia de la imagen previamente adquirida, asociada a la persistencia de cargas eléctricas residuales una vez leído el detector. Esta remanencia exige aplicar técnicas de borrado de la imagen previa algo más complejas que con otros materiales.[\[9\]](#)

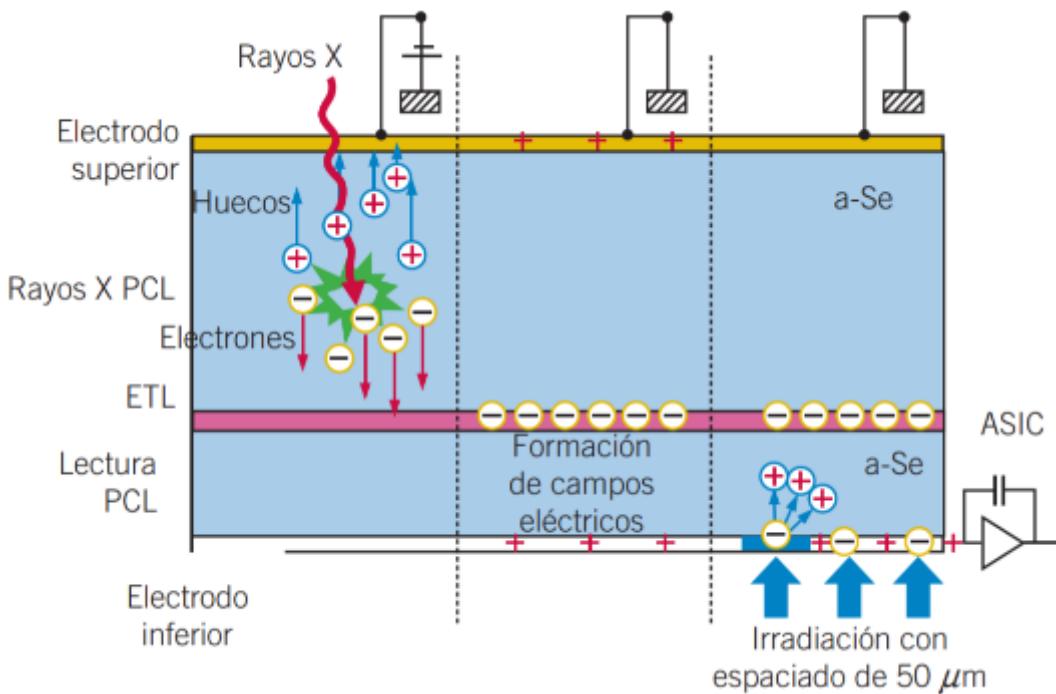


Figura 2.3.8: Esquema del detector de doble capa de a-Se. PCL: Capa fotoconductiva; ETL: Capa de captura de electrones.[9]

2.3.2. Sistema automático de control de exposición

En la década de 1980, el sistema de control automático de exposición (AEC) se implementó en equipos de mamografía con el objetivo de proporcionar una exposición y penetración uniformes y reproducibles de los tejidos mamarios, independientemente de su grosor o composición. Las unidades mamográficas modernas hacen un uso intensivo del AEC para medir la duración de la exposición.

La corriente del tubo a menudo es fija o variada en bandas anchas. Los AEC mamográficos pueden ser muy sofisticados, permitiendo la atenuación del seno, la energía del rayo y seleccionando automáticamente la mejor combinación de objetivo/filtro y kV en los diseños más sofisticados.

Los dispositivos AEC funcionan midiendo la cantidad de radiación que llega al receptor de imagen y terminando la producción de rayos X cuando se obtiene un nivel adecuado de exposición en el detector. Este sistema está compuesto por uno (o más) detectores de radiación, amplificador de señal, selector de densidad, circuito comparador, interruptor de terminación y un temporizador de respaldo. Teniendo en cuenta la configuración más típica en los sistemas de mamografía, los rayos X transmitidos a través del paciente generan instantáneamente una pequeña señal en los sensores AEC ubicados detrás del receptor de imagen. Un amplificador aumenta la señal, que se envía a un comparador de voltaje y un circuito de integración. Cuando la señal acumulada es igual a un valor de referencia preseleccionado, un impulso de salida termina la exposición. Si el detector o el circuito falla, un dispositivo de seguridad de temporizador de respaldo termina la radiografía después de un tiempo preestablecido. Los dispositivos AEC requieren calibración para establecer el nivel adecuado

del detector de referencia para varias condiciones de exposición a rayos X.

En las unidades digitales modernas, el control automático de la exposición se puede configurar para proporcionar una relación constante de contraste/ruido con el aumento del grosor de la mama comprimida, o una configuración de dosis baja comprometida en la que se permite que la relación de contraste/ruido disminuya lentamente al aumentar el espesor de la mama comprimida a cambio de una reducción en dosis para los senos más grandes.

Estas funciones se logran utilizando un conjunto complejo de relaciones entre kVp, combinación de objetivo/filtro y tiempo de exposición, que utilizan entradas de la posición de la paleta de compresión, así como mediciones sensibles a la energía de la transmisión de radiación a través del seno. Algunos diseños utilizan la salida del propio receptor de imagen para proporcionar una señal al control de exposición automático, y esto se puede configurar de varias maneras mediante el uso de software para proporcionar una gama de patrones de sensores, o ubicar automáticamente el área más densa del seno para proporcionar una señal de referencia.

En los sistemas de mamografía digital, los valores de escala de grises no dependen de la exposición incidente sino del procesamiento y visualización de la imagen. La calibración AEC en mamografía digital no utiliza la densidad óptica como se usa en los sistemas de mamografía de película. Hoy en día, los dispositivos AEC están calibrados para adaptarse a la respuesta de energía del receptor de imagen. [8]

2.3.3. Características generales de las imágenes mamográficas digitales

Una mamografía digital está compuesta por un número finito de puntos (o píxeles), donde cada píxel tiene un valor de brillo dictado por un valor numérico almacenado, correspondiente a la atenuación de los rayos X por el tejido mamario. Las imágenes digitales tienen la ventaja de que pueden ser mejoradas y manipuladas por computadora para extraer la máxima cantidad de información de diagnóstico. Las imágenes digitales se pueden almacenar, transferir, copiar sin detrimento y recuperar de una manera muy eficiente utilizando técnicas de almacenamiento masivo de datos por computadora. Tienen la desventaja de un límite a la resolución espacial causada por el tamaño de píxel finito.

Los receptores de mamografía digital generalmente tienen una respuesta lineal entre el valor de píxel y la dosis de radiación incidente en el píxel en un rango dinámico muy amplio, generalmente un factor de unos 10,000: 1. Por lo tanto, la elección de qué dosis de paciente se requiere para la mamografía digital se basa en la relación señal-ruido requerida para una imagen de diagnóstico en lugar de una dosis de radiación específica para el receptor.

2.3.4. Herramientas de visualización

Se espera que las estaciones de trabajo de pantalla proporcionen una interfaz de usuario que proporcione un rendimiento eficiente de imágenes y una gama de herramientas de pantalla que incluyen

típicamente:

- Ampliación, zoom y paneo (roam).
- Ajuste de contraste y brillo (ventaneo).

El sistema visual humano solo es capaz de distinguir unos 100 niveles de gris en una imagen, incluso en condiciones de visualización ideales, por lo tanto, si toda la información presente en una imagen digital se muestra en el monitor de una vez, pequeñas diferencias de contraste, aunque registrado con éxito, no sería distingible. La solución a este problema es mostrar solo un rango seleccionado de valores de píxeles, aumentando así el contraste mostrado para ese subconjunto de niveles. Esta ventana de valores de número de píxel se define por un ancho y un nivel de ventana. Al modificar los ajustes de ancho y nivel de la ventana de visualización, el observador puede optimizar la visualización del rango de niveles de gris para la tarea de diagnóstico que se está realizando, y se puede mostrar cualquier contraste registrado en la imagen, pero el tiempo necesario para realizar muchos de estos ajustes puede convertirse en un factor a la hora de reportar grandes volúmenes de imágenes.

- Giro de imagen y rotación.
- Inversión negro / blanco.
- Medición espacial.
- Mejora de bordes y reducción de ruido (filtrado de frecuencia espacial). Las imágenes se pueden pensar y analizar como conjuntos de frecuencias espaciales. En general, las frecuencias espaciales bajas se asocian con un gris uniforme o gradientes que cambian lentamente, mientras que las frecuencias espaciales altas se asocian con cambios repentinos en el brillo, como en bordes afilados o patrones de puntos o líneas. Al aplicar un filtro de frecuencia espacial, los rangos de frecuencias espaciales se pueden mejorar o atenuar. La mejora de las altas frecuencias espaciales mejora el contraste de los bordes afilados, por ejemplo: microcalcificaciones y estructuras lineales, y en general, agudiza la imagen. La atenuación de altas frecuencias difumina la imagen, y esto puede usarse para reducir la apariencia del ruido cuántico en algunas situaciones. Varias capas de procesamiento de imágenes, incluido el filtrado de frecuencia espacial, se utilizan de forma rutinaria en la mamografía digital.

2.4. Densidad mamaria

El tejido mamario está compuesto principalmente por dos tipos de componentes:

- Tejido adiposo subcutáneo.

- Tejido tejido fibroglandular.

Este último se trata de una mezcla de tejido glandular funcional (parénquima) y tejido conectivo fibroso de sostén (estroma).

La apariencia radiográfica de la mama varía entre las mujeres debido a las diferencias en la composición del tejido y las diferencias en las propiedades de atenuación de los rayos X de los mismos. La grasa es más radiolúcida, es decir que tiene un coeficiente de atenuación menor, que el tejido fibroglandular, por lo tanto, aparece más oscura en la imagen mamográfica. De este modo, las regiones de mayor intensidad de brillo, asociadas al tejido fibroglandular, son referidas como densidad mamográfica. Por tanto, el patrón de brillantez en una imagen mamográfica permitirá inferir la relativa prevalencia de este componente en la mama. Es difícil lograr un buen contraste en mamografías de mujeres con tejido denso extenso, que puede ocultar algunos de los signos de anomalías en las mamas, las cuales generalmente son igualmente radiopacas. En consecuencia, es más difícil diagnosticar el cáncer de mama en estas mujeres. El seno denso se considera un factor importante que contribuye a la interpretación de falsos negativos, en otras palabras, la sensibilidad del estudio disminuye con el aumento de la densidad mamaria.

Independientemente de lo anterior, hay estudios que sugieren que una mayor densidad mamaria también contribuye a un aumento en el riesgo de contraer cáncer. Las mujeres con una proporción mayor al 75 % de tejido fibroglandular tienen un riesgo de padecer cáncer de mama entre cuatro a seis veces mayor que aquel entre las mujeres con poco o ningún tejido denso. [10] [11] [12]

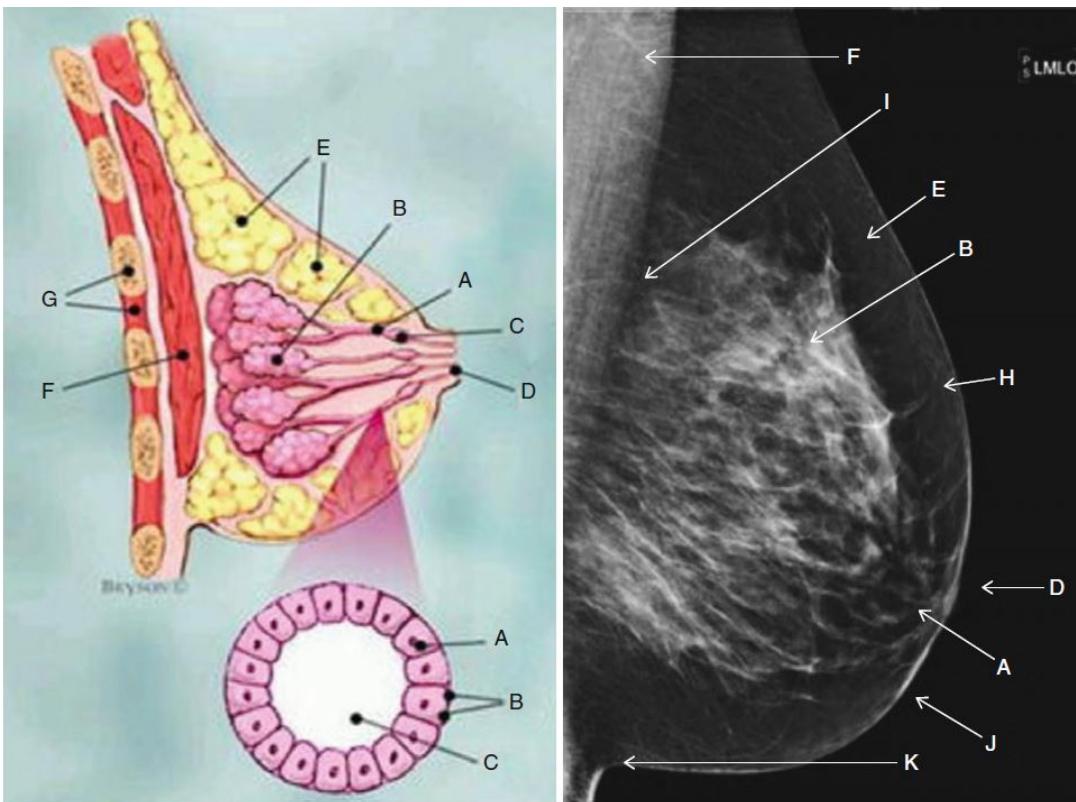


Figura 2.4.1: Anatomía interna de la mama en una imagen radiográfica. A. Conducto galactóforo. B. lóbulos. C. Sección de un conducto galactóforo. D . Pezón. E. Tejido adiposo. F. Músculo pectoral mayor. G. Espacio Retromamario. H. Ligamentos de Cooper. I. Espacio retromamario. J. Piel. K. Folio inframamario.[8]

2.4.1. Factores que influyen en la densidad mamaria

- *Edad:* Un alto porcentaje de mujeres menores de 30 años tiende a tener senos densos, (aproximadamente el 90 % de densidad frente al 10 % de grasa). La tasa de densidad disminuye constantemente en aproximadamente 1–2 % por año. A los 40 años, la relación es de 80/20; a los 50 años 70/30. Es aproximadamente 50/50 a los 65 años. Se trata de una atrofia fisiológica gradual del tejido glandular mamario a partir de finales de la cuarta década de la mujer. Esto se debe a que la función de los ovarios disminuye, lo que hace que los tejidos conjuntivos de soporte en el seno sean reemplazados por tejido adiposo. También se observan una atrofia lobular progresiva y una reducción en el componente glandular con un aumento en el tejido graso.
- *Embarazo y lactancia:* Durante el embarazo, los aumentos de estrógeno, progesterona y prolactina conducen al crecimiento de los acinos, la hiperplasia del epitelio lactogénico (productor de leche) y un aumento de las células mioepiteliales en preparación para la producción de leche. Los lóbulos se agrandan hasta que solo los tabiques fibrosos delgados los separan. Una vez que cesa la lactancia, los senos sufren un grado de involución. Una mujer que tuvo hijos tiene mamas más grasas que una de la misma edad que no tuvo.

- *Estado hormonal:* Los niveles de estrógeno, que disminuyen con la edad y el estado menopáusico, pueden llevar a una disminución de la densidad mamográfica. El estrógeno, post menopausia, se asocia positivamente con el índice de masa corporal.
- *Índice de masa corporal:* Las mujeres con un índice de masa corporal grande (IMC) tienden a tener mamas grandes con tejido graso significativo y con una pérdida asociada en la densidad. El pecho es un almacén para la grasa y como una mujer gana o pierde peso, esto tendrá un efecto en el porcentaje de tejido mamario denso.[8]

2.4.2. Clasificación de la densidad mamaria

Existen diferentes escalas de evaluación cualitativa de la densidad. La clasificación de la composición mamaria por parte de Wolfe [11] es la siguiente:

- *N1:* predominantemente grasa.
- *P1:* prominencia ductal < 25 %.
- *P2:* prominencia ductal > 25 %.
- *DY:* displasia extensa.

Boyd [13] por su parte, clasificó el tejido mamario en seis categorías de acuerdo al porcentaje de tejido fibroglandular:

- *A:* 0 %.
- *B:* >0–10 %.
- *C:* >10–25 %.
- *D:* >25–50 %.
- *E:* >50–75 %.
- *F:* >75 %.

Haremos especial énfasis en la clasificación BI-RADS del Colegio Americano de Radiología (ACR por sus siglas en inglés, American College of Radiology)[5] por ser esta la forma en la que mostraremos los resultados:

- *a:* mamas compuestas por tejido adiposo casi en su totalidad. La mamografía es muy sensible en este contexto, siempre que se incluya en el campo de la imagen el sector que contiene la anormalidad.

- *b*: se observan sectores dispersos de densidad fibroglandular.
- *c*: mamas que presentan densidad heterogénea, lo que puede ocultar algunos nódulos pequeños.
No es poco habitual que algunos sectores de las mamas que tienen estas características sean bastante densos mientras que otros sean mayormente adiposos. Cuando es así, puede resultar de utilidad describir la ubicación del tejido más denso ya que podrían contener pequeñas lesiones no calcificadas ocultas.
- *d*: mamas muy densas, lo que disminuye la sensibilidad mamográfica. Esta categoría corresponde a la sensibilidad mamográfica más baja.

Ha de notarse que el ACR ya no distingue la densidad en intervalos porcentuales al describir las cuatro categorías. Este cambio, hace hincapié en las descripciones de la densidad que reflejan el efecto de enmascaramiento que tiene el tejido fibroglandular denso sobre la representación mamográfica de las lesiones no calcificadas, dado que han llegado a la conclusión de que la asociación entre la densidad mamaria calculada en términos subjetivos y los cambios de la sensibilidad mamográfica tiene más importancia clínica que el efecto relativamente menor que puede tener el porcentaje de densidad mamaria como indicador del riesgo de cáncer.

Algunas mamas pueden parecer más o menos densas en las mamografías digitales que en la mamografía convencional. La mejor representación de la línea cutánea que ofrece la mamografía digital permite calcular con mayor precisión (y, generalmente, en mayor cantidad) el volumen de grasa subcutánea. No obstante, no se ha observado ningún cambio en la distribución de las categorías de la densidad al comparar la mamografía digital con la mamografía convencional.



Figura 2.4.2: Mamografía digital con ACR a.

2

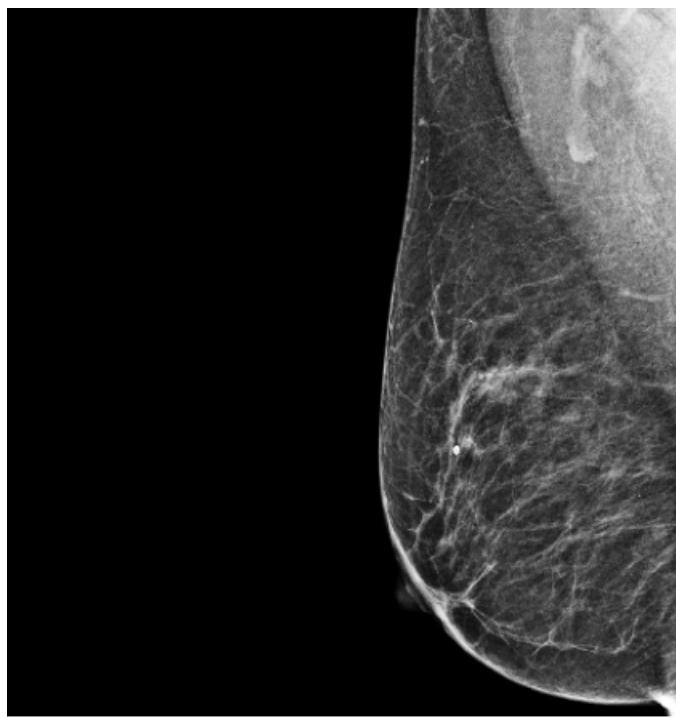


Figura 2.4.3: Mamografía digital con ACR b.

3



Figura 2.4.4: Mamografía digital con ACR c.

4

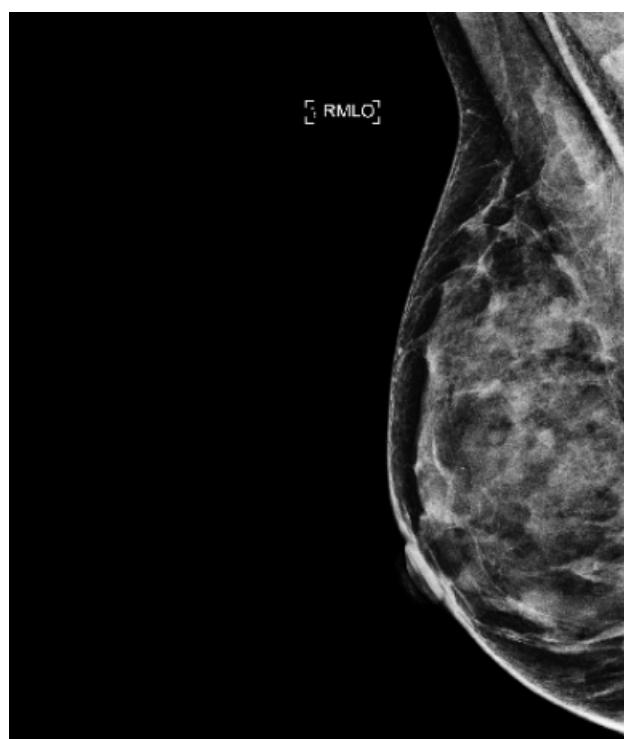


Figura 2.4.5: Mamografía digital con ACR d.

5

2.5. Evaluación diagnóstica de la mama

La mamografía y otros métodos de diagnóstico de la mama, como la ecografía y la resonancia magnética, también sirven para evaluar a las mujeres que presentan algún signo o síntoma sugerente de cáncer de mama. No obstante, no hay ningún estudio ni grupos de estudios que garantice la ausencia de cáncer de mama. El examen físico permite evaluar características tisulares diferentes de las que se observan en la mamografía, por lo que brinda información única respecto de los tejidos que están en estudio. Así como es preciso tomar decisiones en función de los signos sospechosos que presenta la mamografía pese a que la palpación mamaria haya sido normal, también es necesario tener en cuenta los signos clínicos aunque la mamografía sea normal.

Por otra parte, es necesario evaluar por separado, sin tener en cuenta la mamografía, todo signo clínico preocupante que no tenga correlato mamográfico, para lo cual suele ser de utilidad la ecografía. Se ha demostrado que, cuando la combinación de mamografía y ecografía tiene resultados negativos, la probabilidad de cáncer se ubica entre el 0,1 % Y el 4 %.

Está comprobado que la mamografía negativa no descarta el diagnóstico de malignidad y que es necesario biopsiar las regiones que presentan signos clínicos sospechosos aunque la mamografía sea negativa.

Pese a que cabe indicar una biopsia si se encuentra una alteración palpable sospechosa, la mamografía sigue siendo importante para evaluar esa región y, a la vez, para estudiar el resto del tejido de la mama homolateral y la mama contralateral con el fin de detectar tumores malignos asintomáticos. Asimismo, es importante que las mujeres y su médico entiendan que el tamizaje mamográfico no es un método perfecto, por lo que es importante que el médico preste atención a todo cambio mamario que esté desvinculado del ciclo menstrual, independientemente del tiempo que haya transcurrido después de que la mamografía y la palpación mamaria hayan tenido resultado negativo. [5]

2.6. Estándar DICOM

DICOM (Digital Imaging and Communication in Medicine) es el estándar para el intercambio de imágenes médicas, pensado para su manejo, visualización, almacenamiento, impresión y transmisión. Incluye la definición de un formato de fichero y de un protocolo de comunicación de red.

El estándar describe el formato de archivos y la especificación de los datos primordiales de un paciente en la imagen así como el encabezado requerido, describiendo un lenguaje común a distintos sistemas médicos. De esta forma las imágenes vienen acompañadas de mediciones, cálculos e información descriptiva relevante para diagnósticos. [14]

Un elemento de DICOM que es particularmente importante desde el punto de vista de los informes radiológicos es la función de visualización estándar en escala de grises (GSDF). Esto se basa en un modelo psicofísico del sistema visual humano y está diseñado para maximizar el número de

diferencias notables que puede reproducir una pantalla determinada, y para proporcionar una escala de grises perceptualmente lineal, con el mismo pequeño cambio de contraste visible en una parte oscura de la imagen como en una parte clara. Si la GSDF se implementa correctamente para un monitor determinado, debe proporcionar la mejor visualización que el monitor es capaz de realizar en las condiciones de visualización donde se usa.

2.7. Procesamiento digital de imágenes

El procesamiento digital de imágenes es el conjunto de técnicas que se aplican a las imágenes digitales con el objetivo de mejorar su calidad o facilitar la búsqueda de información.

Una imagen digital puede ser representada por una función $f(x, y)$ donde x e y denotan coordenadas espaciales, y el valor de f (intensidad o luminancia) en cualquier punto (x, y) es proporcional al nivel de gris (brillo) de la imagen en ese punto. Así mismo puede ser convenientemente representada por una matriz I de tamaño $M \times N$ de la forma:

$$\begin{matrix} I(1, 1) & I(1, 2) & \dots & I(1, N) \\ I(2, 1) & I(2, 2) & \dots & I(2, N) \\ I = & \ddots & \ddots & \ddots \\ & \vdots & \vdots & \vdots \\ I(M, 1) & I(M, 2) & \dots & I(M, N) \end{matrix}$$

Donde cada elemento se denomina píxel. Los cuales, en una imagen monocromática típica, son del orden de 28 o 256 niveles de gris.

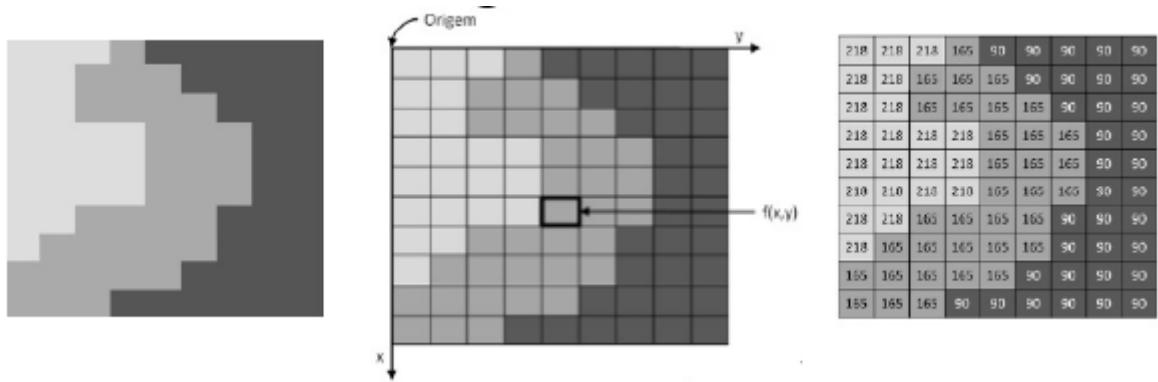


Figura 2.7.1: Definición de imagen.

Ya que hablamos sobre lo que representa un píxel, explicaremos conceptos fundamentales que tienen que ver con la relación entre ellos.

2.7.1. Relaciones básicas entre píxeles

Las relaciones básicas entre píxeles son [15]:

- *Vecindad*: un píxel posee vecinos verticales, horizontales y diagonales. Al conjunto se lo denomina 8 vecinos.
- *Contorno*: el contorno de orden 8 de un píxel particular, estará formado por todos los píxeles que lo rodean.
- *Conectividad*: son criterios de proximidad (distancia) o bien criterios de similitud en el nivel de gris. (4-conectividad u 8-conectividad).
- *Adyacencia*: dos píxeles son adyacentes si presentan algún tipo de relación por conectividad.

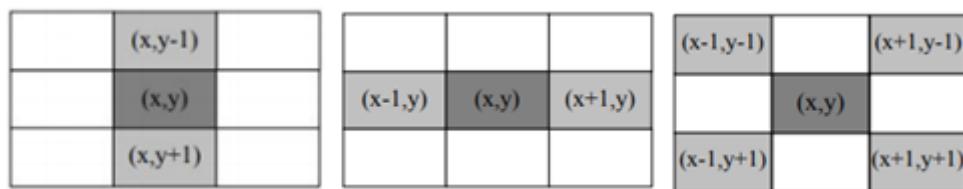


Figura 2.7.2: Relaciones entre píxeles. [15]

2.7.2. Etapas fundamentales del procesamiento digital de imágenes

Los pasos principales que tiene que seguir un sistema de captación de imágenes para obtener la imagen y luego poder interpretarla, se resumen en el siguiente esquema [15]:

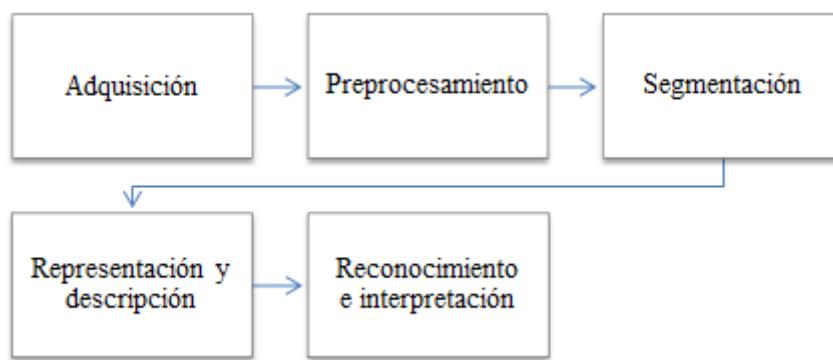


Figura 2.7.3: Etapas fundamentales del procesamiento digital de imágenes

Dado que la adquisición de una imagen de mamografía digital ha sido explicada anteriormente, nos enfocaremos en las etapas que le siguen a ella.

2.8. Preprocesamiento

Las técnicas de procesado pretenden mejorar o realzar las propiedades de la imagen para facilitar las siguientes operaciones de la Visión Artificial, tales como las etapas de segmentación, extracción de las características y finalmente la interpretación automática de las imágenes.

La visión artificial o visión por ordenador es una disciplina científica que incluye métodos para adquirir, procesar, analizar y comprender las imágenes del mundo real con el fin de producir información numérica o simbólica para que puedan ser tratados por un ordenador. Tal y como los humanos usamos nuestros ojos y cerebros para comprender el mundo que nos rodea, la visión artificial trata de producir el mismo efecto para que los ordenadores puedan percibir y comprender una imagen o secuencia de imágenes y actuar según convenga en una determinada situación. [16]

Las técnicas de preprocesamiento son operaciones para preparar la imagen. Para ello se utilizan [15]:

- *Técnicas de mejora:* Se utilizan el histograma, transformaciones y filtros para quitar ruidos, mejorar el contraste o enfatizar alguna característica relevante.
- *Técnicas de restauración:* a partir del conocimiento del proceso de formación de la imagen se pretende obtener la imagen sin deformación alguna. Esto ocurre cuando una cámara tiene mal la lente e introduce distorsiones.

2.8.1. Operaciones de punto

Las operaciones de punto son operaciones de memoria cero (ZMO), es decir, dada una imagen almacenada en memoria, no necesitan de un espacio adicional para realizar la transformación de la imagen. Las operaciones de píxel tienen como requisito conocer en cada paso el valor de la luminancia (nivel de gris) del píxel al que se le aplica la transformación. Una vez que se ha realizado la transformación de un valor de luminancia, éste no vuelve a ser necesario en todo el algoritmo, por eso a este tipo de operaciones se las conoce como operaciones de memoria cero.

El mapa de transición de luminancias indica cómo se produce la transformación entre el anterior nivel de luminancia y el siguiente. Es decir, muestra la gráfica de la función de transformación que se aplica a la imagen (o la región de la misma a manipular).[15]

Complementación (negativo)

Consiste en sustituir el valor de la luminancia en cada punto por el de su complementario.

Si llamamos a [15]:

L : nuevo valor de luminancia y $u(y, x)$: luminancia de imagen original, ambos mayores a cero.

Se logra un efecto de negativo digital de la imagen cuando $L = u_{max}(y, x) - u(y, x)$

Binarización

Consiste en generar una imagen en dos tonos (generalmente blanco y negro) a partir de otra con múltiples niveles de gris.

Como refleja el mapa de transición de luminancias, el valor umbral m determina qué valor pasan a tener los píxeles (blanco o negro). Generalmente el fondo es negro y el objeto, en primer plano, blanco. [15]

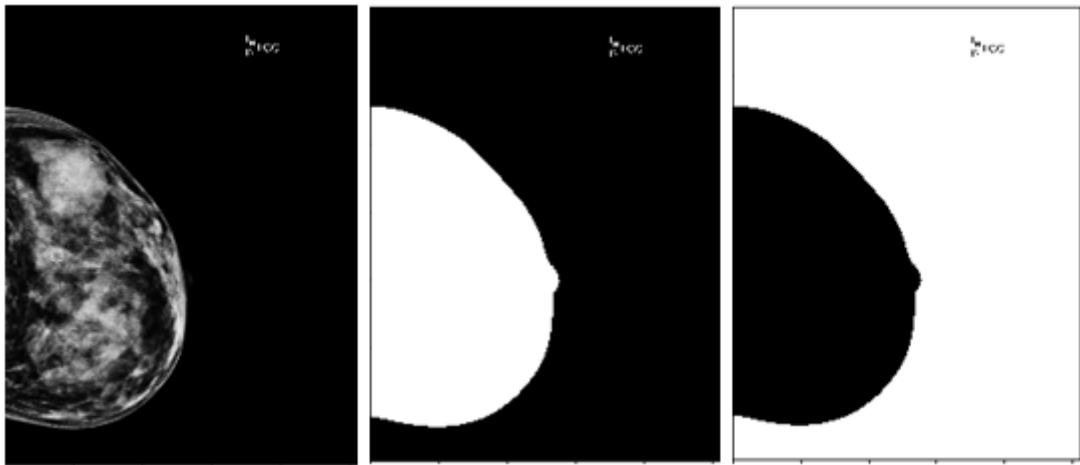


Figura 2.8.1: Binarización y complementación respectivamente.

Suma

Hay dos formas de realizar una suma [15]:

- Suma entre dos imágenes $L = u(y, x) + v(y, x)$
- Suma de una imagen con una constante $L = u(y, x) + k$

Resta

Al igual que la suma, se puede restar [15]:

- Dos imágenes $L = u(y, x) - v(y, x)$
- Una imagen y una constante (decrementando los niveles de grises) $L = u(y, x) - k$

Multiplicación

Esta operación se puede realizar de dos modos [15]:

- Multiplicando los píxeles de una imagen por una constante: se realiza un escalado en los niveles de gris: $L = u(y, x) \times k$
- Multiplicando los píxeles de una imagen por los de otra, píxel a píxel: $L = u(y, x) \times v(y, x)$

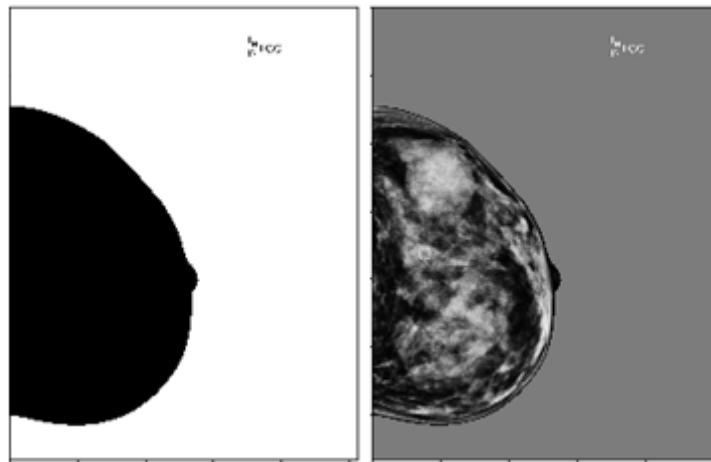


Figura 2.8.2: Imagen negativa * 2000 + imagen original.

2.8.2. Operaciones morfológicas

Estas operaciones realizan transformaciones en las formas de los objetos de la imagen [15].

Dilatación

La dilatación no es una operación de memoria cero. Se necesita conocer información de la imagen original.

Se aplica típicamente a imágenes binarias y su efecto básico es agrandar gradualmente los límites de una región de píxeles de primer plano (agranda la imagen, haciendo más grandes sus contornos).

Erosión

Al igual que la dilatación también se aplica típicamente a imágenes binarias. Realiza una erosión de los límites de una región del frente.

Apertura

De forma parecida a la erosión, pretende eliminar algunas zonas de los píxeles del frente, siendo en general menos destructiva que la erosión. Se utiliza para eliminar ruido de las imágenes. Es una combinación de las dos operaciones morfológicas básicas:

$$\text{Apertura} = \text{erosión} + \text{dilatación}$$

Se superpone el kernel por los píxeles del primer plano de la imagen y se mantienen aquellas regiones que tengan una forma similar a este kernel, o que puedan contener completamente al kernel, mientras se eliminan las otras regiones de los píxeles del frente.

Cierre

Es parecido a la dilatación ya que ambas tienden a agrandar los bordes de regiones del frente, pero más constructiva que esta operación. El cerramiento dilata suavemente, cerrando un poco los huecos. Es también una combinación de las operaciones morfológicas básicas:

$$\text{Cierre} = \text{dilatación} + \text{erosión}$$

Se utiliza un kernel y se superpone por los píxeles del fondo (excepto en los bordes), preservando el fondo de regiones que tienen una forma similar a él o que pueden contener completamente al kernel, mientras que se eliminan todas las otras regiones de píxeles del fondo donde no entre (poniendo su valor a 1).



Figura 2.8.3: Imagen binaria, dilatada y erosionada.

2.8.3. Filtrado

El filtrado permite realizar o atenuar, suavizando los detalles, alguna característica de la imagen, así como eliminar el ruido o las partes que no interesen. [15]

Es necesario aplicar convolución: Imagen * Filtro

$$g(y, x) = f(y, x) * h(y, x)$$

Siendo: $g(y, x)$: Imagen Filtrada. $f(y, x)$: Imagen Original. $h(y, x)$: Operador de Convolución.

El operador de convolución para realizar el filtrado, es un kernel, el cual estará formado por una matriz que se aplica a todos los píxeles de la imagen. Éste se centra en el píxel a tratar y luego se calcula el producto punto entre él y la porción de imagen superpuesta, como vemos en la imagen a continuación.

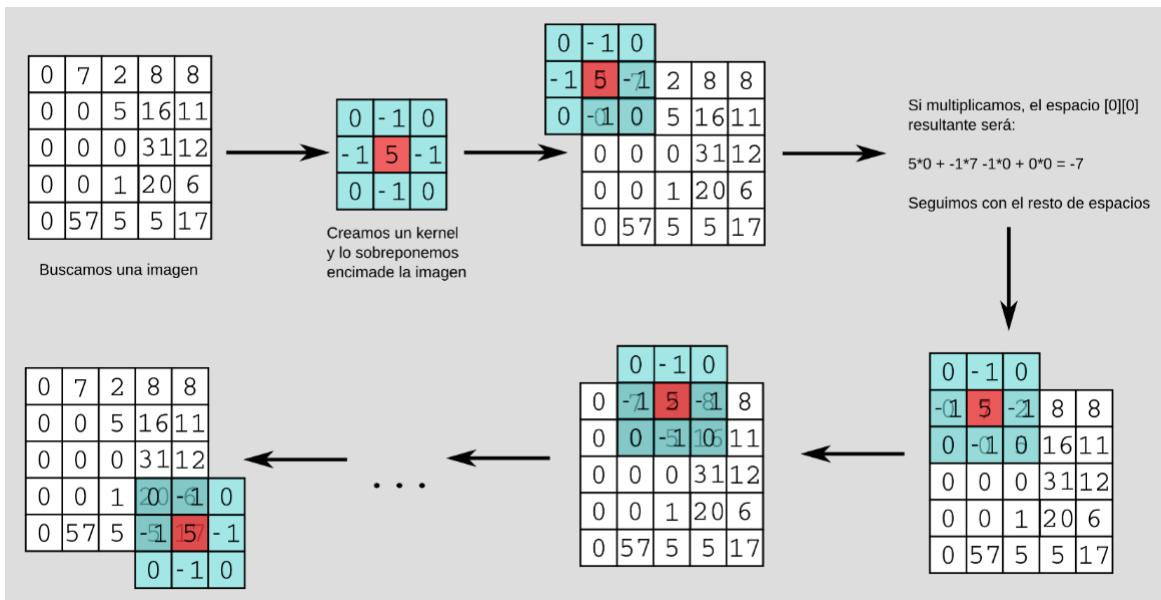


Figura 2.8.4: Convolución. [17]

Si $K \times L$ es el tamaño de la matriz de entrada y $M \times N$ la máscara, la dimensión de la imagen de salida es $(K+M-1) \times (N+L-1)$. Este efecto se debe a cuando el operador convolución pasa por los bordes de la imagen. En los píxeles de los bordes cuando se pivota la máscara, los vecinos del píxel que no existen, por defecto, son considerados nulos. En los algoritmos que implementan esta operación el efecto de los bordes son tratados de tres maneras distintas:

- Se considera que los píxeles vecinos no existentes bien son cero, un valor dado por el usuario o se toma un valor intermedio de los que existen. La imagen de salida es $(K+M-1) \times (L+N-1)$.
- Se hace la misma convolución que en el apartado anterior pero eliminando las filas y columnas exteriores. La imagen de salida es de igual tamaño que la de entrada, $(K \times L)$.
- La convolución se hace sólo en la parte central de la imagen de entrada, esto es, se hace una submatriz tal que al convolucionar todos los píxeles de éstas tienen vecinos conocidos. El tamaño de la imagen de salida es $(K-M+1) \times (L-N+1)$.

Realce de bordes con gradiente direccional

Empleado para destacar y resaltar con mayor precisión los bordes que se localizan en una dirección determinada.

Para ello, este filtro produce un nuevo píxel resaltado si existe una pendiente positiva en la dirección del kernel de convolución. La intensidad del nuevo píxel es directamente proporcional a la intensidad de la pendiente píxel a píxel que el kernel experimente.

Los kernels que se utilizan son los siguientes:

$$\begin{array}{cccc}
 1 & 1 & 1 & -1 & -1 & -1 \\
 1 & -2 & 1 & 1 & -2 & 1 \\
 -1 & -1 & -1 & 1 & 1 & 1 \\
 \text{norte} & & \text{sur} & & \text{este} & \text{oeste}
 \end{array}
 \quad
 \begin{array}{cccc}
 -1 & 1 & 1 & 1 & -1 & -1 \\
 -1 & -2 & 1 & 1 & -2 & -1 \\
 1 & 1 & 1 & 1 & -1 & -1 \\
 \text{sureste} & & \text{noroeste} & & \text{suroeste} & \text{noreste}
 \end{array}$$

Figura 2.8.5: Kernels utilizados en técnica de realce de bordes por medio del gradiente direccional.[18]

Como vemos en la Figura 2.8.5, la suma de los coeficientes es igual a 0. Luego, a medida que la máscara pasa sobre regiones de la imagen con valores de brillo constantes, el resultado es 0, indicando pendiente de brillo nula. En aquellas regiones donde el gradiente genera resultados negativos, el valor de salida se establece también igual a 0 debido a que brillos negativos no están definidos.

Por último, la imagen de gradiente aparece negra toda vez que los valores de brillo son constantes y sólo aquellos bordes con la orientación direccional correcta en la imagen original, aparecen como blancos.[18]

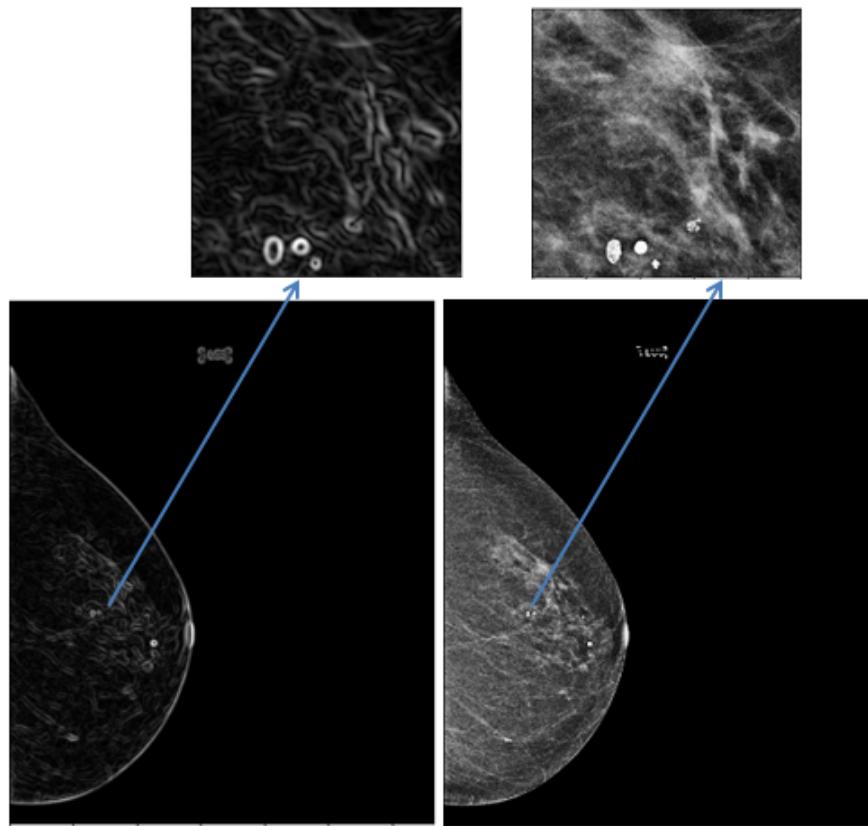


Figura 2.8.6: Imagen resultante luego de aplicar filtro de gradiente.

Filtro gaussiano

El filtro Gaussiano realiza una media ponderada donde los pesos toman la forma de una campana de Gauss. El valor máximo aparece en el píxel central y disminuye hacia los extremos más rápido cuanto menor sea el parámetro de desviación típica s (o σ). El resultado será un conjunto de valores entre 0 y 1. Para transformar la matriz a una matriz de números enteros se divide toda la matriz por el menor de los valores obtenidos [19]. La ecuación para calcularla es:

$$g(x, y) = e^{-\frac{x^2+y^2}{s^2}}$$

$$G(x, y) = \frac{g(x, y)}{\min_{x,y}(g(x, y))}$$

Este tipo de filtro se utiliza fundamentalmente para el suavizado de imágenes con presencia de ruido. El efecto que produce es similar a un desenfoque. La máscara mayormente empleada[19] es la que se muestra en la Figura 2.8.7.

$$W = \frac{1}{16} * \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

Figura 2.8.7: Ejemplo de máscara para filtro gaussiano.[19]

2.9. Segmentación

Consiste en subdividir una imagen en sus partes u objetos constituyentes.[15]

Se utilizan algoritmos de segmentación, que se basan en:

- *Discontinuidades del nivel de gris:* Consiste en dividir una imagen basándose en los cambios bruscos de nivel de gris.
 - *Similaridad del nivel de gris:* Se obtiene la segmentación basándose en las propiedades de distribución de los píxeles, como la intensidad o el color, o en encontrar directamente las regiones.
- Las principales técnicas son:

- Umbralización.
- Crecimiento de regiones.

2.9.1. Umbralización

Hay dos posibles situaciones [15]:

- *Umbral único:* hay dos modos o agrupaciones dominantes de niveles de gris en el histograma (objetos luminosos y fondo oscuro). Para extraer los objetos del fondo hay que elegir un umbral T . Cualquier punto (x, y) para el que $f(y, x) > T$ será un punto del objeto.
- *Umbral multinivel:* hay más de dos modos dominantes (tres en el ejemplo: objetos luminosos, objetos menos luminosos y fondo oscuro). Para separar los objetos entre sí y del fondo hay que elegir los umbrales T_1 y T_2 . Un punto pertenece a una clase si $T_1 < f(y, x) \leq T_2$, a la otra clase si $T_2 < f(y, x)$ y al fondo si $f(y, x) \leq T_1$.

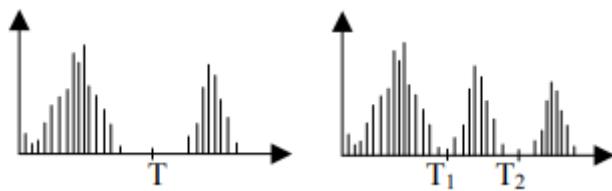


Figura 2.9.1: Umbral único y umbral multinivel.[15]

Expicaremos dos métodos utilizados para encontrar el valor del umbral, de manera tal que nos permita segmentar correctamente la imagen en dos o más regiones.

Método Ridler y Calvard

Es una técnica iterativa para encontrar un umbral, cuyo algoritmo es el siguiente [20]:

1. El histograma es inicialmente segmentado en dos partes tomando como umbral inicial (T) el promedio del rango máximo de niveles de grises.
2. Se partitiona la imagen en dos grupos $R1$ y $R2$ utilizando T .
3. Se calcula los valores medios de las dos regiones segmentadas en el punto anterior ($M1, M2$).
4. Se calcula el nuevo umbral $T = \frac{(M1+M2)}{2}$.
5. Se repiten los pasos 2-4 hasta que el umbral no cambie.

Método de Otsu

Se realiza un procedimiento que selecciona el umbral óptimo maximizando la varianza entre clases mediante una búsqueda exhaustiva.

Partimos de una imagen en niveles de gris con N píxeles y L posibles diferentes niveles. La probabilidad de ocurrencia del nivel de gris i en la imagen es [21]:

$$p_i = \frac{f_i}{N}$$

Siendo f_i la frecuencia de ocurrencia del nivel i -ésimo, con $i = 1, 2, \dots, L$. En el caso particular de umbralización en dos niveles (binarización), los píxeles se dividen en dos clases ($C1$ y $C2$), con niveles $[1, 2, \dots, t]$ y $[t + 1, t + 2, \dots, L]$ respectivamente, donde sus distribuciones de probabilidad son:

$$C1 : \quad \frac{p_1}{w_1(t)}, \dots, \frac{p_t}{w_1(t)} \quad w_1(t) = \sum_{i=1}^t p_i$$

$$C2 : \quad \frac{p_{t+1}}{w_2(t)}, \dots, \frac{p_L}{w_2(t)} \quad w_2(t) = \sum_{i=t+1}^L p_i$$

Las medias para cada una de las clases se definen como:

$$\mu_1 = \sum_{i=1}^t \frac{p_i \cdot p_i}{w_1(t)}$$

$$\mu_2 = \sum_{i=t+1}^L \frac{p_i \cdot p_i}{w_2(t)}$$

2.10. Representación y descripción

Una vez que una imagen está segmentada en diferentes regiones, es habitual representarla y después describir esa representación de una forma que facilite su posterior procesado mediante el ordenador. [15]

Esta etapa puede realizarse con dos enfoques distintos:

- *En términos de sus características externas (su contorno):* características de forma.

- En términos de sus características internas (los píxeles que comprenden la región): color y textura (propiedades de reflectividad).

Hay muchos métodos que permiten obtener características significativas que sean capaces de representar un objeto o una imagen particular. Procederemos a explicar los utilizados en este proyecto.

2.10.1. Área

El cálculo del área de la región de interés(ROI) es algo fácil de realizar, y si bien nada tiene que ver con la textura, de cierta forma se relaciona con el nivel de gris de esta región, ya que se necesita como primera medida separarla del fondo (cuyo nivel de gris es 0 o muy bajo), para luego obtener el resultado dimensional.

2.10.2. Operador gradiente

El gradiente de una función escalar multivariable como por ejemplo, $f(x, y)$, denotado como ∇f empaqueta toda la información de sus derivadas parciales en un vector [15]:

$$\nabla f(x, y) = [G_x(x, y) - G_y(x, y)] = \left[\frac{\partial f(x, y)}{\partial x}, \quad \frac{\partial f(x, y)}{\partial y} \right]$$

Como la derivada de una función hace referencia a la pendiente de la misma en un punto, el gradiente se traduciría de la siguiente manera.

Si nos posicionamos en un punto (x_0, y_0) , en el espacio de entrada de f , el vector $\nabla f(x_0, y_0)$ indica en qué dirección se incrementa el valor de f lo más rápido posible. Siendo f en nuestro caso el nivel de gris del píxel.

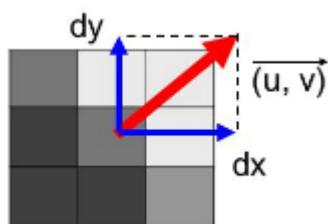


Figura 2.10.1: $\nabla f(x_0, y_0) = (u, v)$ establece la dirección de máximo crecimiento de nivel de gris.

Al ser el gradiente un vector (posee módulo, dirección y sentido) hay dos valores a tener en cuenta:

- El módulo es igual a la máxima variación de $f(x, y)$ por unidad de distancia en la dirección del gradiente.

$$|\nabla f(x, y)| = \left[G_x(x, y)^2 - G_y(x, y)^2 \right]^{1/2}$$

- La fase indica la dirección de la máxima variación.

$$\alpha(x, y) = \arctan\left(\frac{G_y(x, y)}{G_x(x, y)}\right)$$

Este operador es muy utilizado en el proyecto para obtener características de texturas al detectar bordes dentro de la región de interés.

2.10.3. Histograma

De forma cuantitativa, el histograma de una imagen consiste en el número de píxeles que tiene la imagen para cada uno de los niveles de gris. Se representa mediante un eje de coordenadas cartesianas. En las abscisas se representan los niveles de gris y en ordenadas, el número de píxeles.[15]

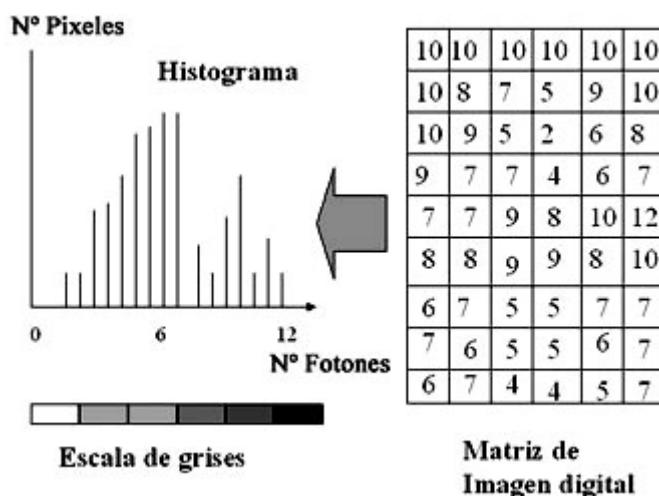


Figura 2.10.2: Representación del histograma de una imagen digital monocromática.

A su vez el histograma sirve para corroborar los valores estadísticos calculados a partir de la matriz de imagen digital (cuyos elementos corresponden a los distintos niveles de grises).

Estos valores representan características que pueden utilizarse para el reconocimiento de una imagen particular. Ellos son [22]:

- *Media Aritmética:* Informa sobre la tendencia general de la variable X en una muestra de N sujetos.

$$\bar{X} = \frac{\sum x_i}{N}$$

- *Mediana:* Puntuación en X que divide la distribución en dos partes iguales, es decir, deja por debajo y por encima de sí al 50 % de las observaciones.

Si N es par: valor central.

Si N es impar: media aritmética de valores centrales $\frac{(Md_{n_1} + Md_{n_2})}{2}$

- *Moda:* Valor de la variable X que más aparece en nuestros datos.

- **Varianza:** Es el promedio de las distancias al cuadrado desde los valores de X hasta la media \bar{X} en una muestra de N sujetos.

$$S_x^2 = \frac{\sum (X_i - \bar{X})^2}{N}$$

- **Desviación estándar:** Es una medida de la dispersión de las observaciones a la media, un promedio de la distancia de las observaciones a la media.

$$S_x = \sqrt{S_x^2}$$

- **Asimetría:** La asimetría de una distribución hace referencia al grado en que los datos se encienden por encima y por debajo de la tendencia central.

$$A_s = \frac{\sum x_i^3}{N S_x^3}$$

Donde: $x_i^3 = (X_i - \bar{X})^3$

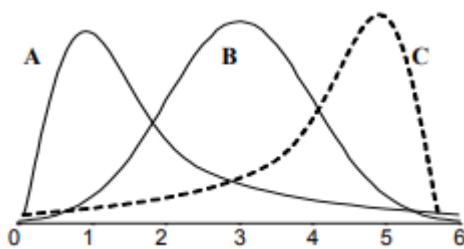


Figura 2.10.3: Asimetría positiva ($A_s > 0$). B) Simetría($A_s = 0$). C) Asimetría negativa ($A_s < 0$). A) Asimetría positiva ($A_s > 0$). B) Simetría($A_s = 0$). C) Asimetría negativa ($A_s < 0$).

- **Curtosis:** La curtosis hace referencia al grado de apuntamiento de una distribución.

$$C_r = \frac{\sum x_i^4}{N S_x^4}$$

Donde: $x_i^4 = (X_i - \bar{X})^4$

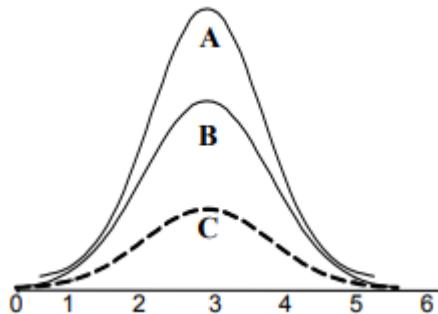


Figura 2.10.4: Distribución leptocúrtica ($C_r > 0$). B) Distribución mesocúrtica ($C_r = 0$). C) Distribución platicúrtica ($C_r < 0$).

2.10.4. Entropía de permutaciones

El concepto de entropía se utiliza para cuantificar la regularidad de una serie temporal, de manera que, cuanto más regular es una serie, más predecible y menos compleja será. Por tanto, en cualquier serie temporal, que representa una variable de salida del sistema, la entropía es una medida de su incertidumbre.[\[23\]](#)

Supongamos que un evento (variable aleatoria) tiene un grado de indeterminación inicial igual a K (existen K estados posibles), la probabilidad de que la variable tome el valor del primer estado posible (K_1) se define:

$$P_{K_1} = \frac{n}{N}$$

Siendo n la cantidad de veces que la variable toma el valor de K_1 en N muestras.

La entropía se define:

$$H = - \sum_{i=1}^K p_i \log_2 (p_i)$$

Esto quiere decir que si una serie temporal es constante (siempre toma el mismo valor), la probabilidad de que tome ese valor es 1 y por lo tanto $\log_2(1)$ y su entropía será 0.

En una imagen monocromática $H = 0$ si los píxeles toman un solo valor constante, por ejemplo, 0 (imagen totalmente negra) o 1 (imagen totalmente blanca).

Un punto importante es que las medidas clásicas de entropía, como la entropía de Shannon descuidan las relaciones temporales entre los valores de la serie temporal, por lo que la estructura y los posibles patrones temporales presentes en el proceso no se contabilizan. Si por ejemplo dos series de temporales son $X_1 = \{0, 0, 1, 1\}$ y $X_2 = \{0, 1, 0, 1\}$ se cumple que $H(X_1) = H(X_2)$.

Una forma de abordar este inconveniente es mediante la comparación de los valores vecinos en una serie de tiempo, en lo que se basa la entropía de permutaciones.

Para facilitar la comprensión, el concepto de entropía de permutaciones se abordará a partir de un ejemplo:

Si tomamos la serie temporal $X(t) = (1, 7; 2, 1; 1, 5; 1, 4; 2)$ y definimos los parámetros $d = 3$ (dimensión de inmersión) y $\tau = 1$ (tiempo de retardo), luego del mapeo de la señal obtendremos los vectores $Y_{(d,\tau)}^t$ correspondientes a la serie :

$$Y_{(3,1)}^1 = (1, 7; 2, 1; 1, 5) \quad Y_{(3,1)}^2 = (2, 1; 1, 5; 1, 4) \quad Y_{(3,1)}^3 = (1, 5; 1, 4; 2)$$

Luego los elementos de estos vectores se ordenan de forma ascendente y se define el vector de permutaciones $\Pi_{(d,\tau)}^t$, cuyas componentes son las posiciones de los valores ordenados de $Y_{(d,\tau)}^t$. En nuestro ejemplo estos serían:

$$\Pi_{(3,1)}^1 = (2; 0; 1) \quad \Pi_{(3,2)}^2 = (2; 1; 0) \quad \Pi_{(3,2)}^3 = (1; 0; 2)$$

Cada uno de estos vectores representa un patrón (o forma). Existen $d!$ (factorial del valor de la dimensión de inmersión) posibles patrones (6 en este caso).

Para una secuencia lo suficientemente grande en comparación con $d!$, es posible calcular las frecuencias de ocurrencia de cualquiera de los $d!$ posibles vectores de permutaciones. A partir de estas frecuencias, se puede estimar la entropía de Shannon asociada con las distribuciones de probabilidad de los vectores de permutaciones.

Algunos de los posibles patrones pueden no ocurrir y se los denomina patrones prohibidos.

Si se denota la probabilidad de ocurrencia del patrón i ésimo como $p_i = p(\Pi^i)$, siendo $i < d!$ entonces la entropía de permutaciones normalizada asociada a la serie temporal $X(t)$ es:

$$h_{EP}(X) = \frac{-\sum_{i=1}^{d!} p_i \log(p_i)}{\log d!}$$

2.10.5. Análisis fractal

La geometría euclíadiana, con sus planos y superficies bien definidos y matemáticamente manejables, por lo general sólo se encuentra como una aproximación sobre un estrecho rango de dimensiones. En cambio, es común que los fenómenos y objetos de la vida real muestren propiedades fractales.

La dimensión fractal es la velocidad a la que el perímetro (o área de superficie en tres dimensiones) de un objeto aumenta a medida que se reduce la escala de medición. Intenta condensar todos los detalles de la forma del límite en un solo número que describe la rugosidad de una manera particular.

Desde el punto de vista del análisis de imágenes, las mediciones de dimensiones fractales se pueden usar para estimar y cuantificar la complejidad de la forma o textura de los objetos. La geometría fractal involucra varios enfoques para definir dimensiones fractales, donde la más común es la dimensión de Hausdorff.[\[24\]](#)

Considerando un objeto que posee una dimensión Euclíadiana E , la dimensión fractal D_0 de Hausdorff puede ser calculada por la siguiente expresión:

$$D_0 = \lim_{\epsilon \rightarrow 0} \frac{\log N(\epsilon)}{\log \epsilon^{-1}}$$

Donde $N(\epsilon)$ es el conteo de hipercubos de dimensión E y largo ϵ que van a llenar el objeto.

Si consideramos un objeto representado por una imagen binaria I_b , se puede obtener una aproximación D para D_0 a través del algoritmo de conteo de recuadros (Box Counting Algorithm).

Para el caso de un espacio bidimensional, el algoritmo se puede describir de la siguiente manera:

1. Se divide la imagen en una cuadrícula compuesta de cuadrados de tamaño $(\epsilon \times \epsilon)$.
2. Contar el número $N(\epsilon)$ de cuadrados de tamaño que contiene al menos un píxel del objeto.
3. Se hace variar el valor de ϵ , y se repite el paso 2 sucesivamente, guardando los valores de: $\log N(\epsilon)$ y $\log \epsilon^{-1}$.

4. Finalmente, se aproxima la curva $\log N(\epsilon)$ vs $\log \epsilon^{-1}$ mediante una línea recta utilizando un método de ajuste de línea (por ejemplo, regresión lineal). La dimensión fractal D corresponde a la pendiente de esta línea.

2.10.6. Descriptores de textura de Haralick

Los descriptores de Haralick son un conjunto de medidas de textura basadas en la matriz de co-ocurrencia (o de niveles de gris). Son de naturaleza estadística y para su cálculo, es necesario asumir que la totalidad de la información textural de una imagen está contenida en las relaciones espaciales que se dan entre los distintos niveles de gris de un objeto. Esas relaciones están especificadas en la matriz de co-ocurrencia espacial que son computadas en una dirección específica, o bien para todas: 0° , 45° , 90° y 135° , entre los píxeles vecinos dentro de una ventana móvil dentro en la imagen.

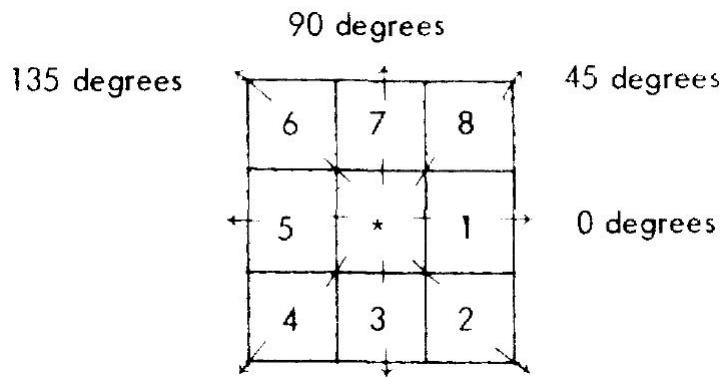


Figura 2.10.5: Celda de resolución.[25]

La matriz de co-ocurrencia describe la frecuencia de un nivel de gris que aparece en una relación espacial específica con otro valor de gris (píxel de referencia y píxel vecino), dentro del área de una ventana determinada. Es un resumen de la forma en que los valores de los píxeles ocurren al lado de otro valor en una pequeña ventana. Las principales propiedades de una matriz de co-ocurrencia son:

- Cuadrada: El rango de los niveles de gris de los píxeles de referencia y el de los vecinos es el mismo, por lo tanto las filas y las columnas tienen idéntico número.
- Tiene el mismo número de filas y columnas que el número de bits de la imagen.
- Es simétrica con respecto a la diagonal.

Con respecto a la matriz de co-ocurrencia simétrica y normalizada hay algunos aspectos a resaltar:

- Los elementos de la diagonal representan pares de píxeles que no tienen diferencias en su nivel de gris. Si estos elementos tienen probabilidades grandes, entonces la imagen no muestra mucho contraste, la mayoría de los píxeles son idénticos a sus vecinos.

- Sumando los valores de la diagonal tenemos la probabilidad que un píxel tenga el mismo nivel de gris que su vecino.
- Las líneas paralelas a la diagonal separadas una celda, representan los pares de píxeles con una diferencia de un nivel de gris. De la misma manera sumando los elementos separados dos celdas de la diagonal, tenemos los pares de pixels con dos valores de grises de diferencia. A medida que nos alejamos de la diagonal la diferencia entre niveles de grises es mayor.
- Sumando los valores de estas diagonales paralelas obtenemos la probabilidad que un píxel tenga 1, 2, 3, etc. niveles de grises de diferencia con su vecino.

Este método permite extraer una gran cantidad de información de textura de imagen por la gran variedad de descriptores que es posible obtener de esta matriz, que hacen posible caracterizar con un conjunto de valores cuantificables cada imagen analizada. Las características propuestas por Haralick [25] son las siguientes:

Notación:

$p(i, j)$ es la (i, j) -ésima entrada en una matriz de dependencia espacial de tonos grises normalizada. $p(i, j) = \frac{P(i, j)}{R}$

$p_x(i)$ es la entrada en la matriz de probabilidad marginal obtenida sumando las filas de $p(i, j)$

$$p_{x+y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j), \quad \text{con } i + j = k = 2, 3, \dots, 2N_g$$

$$p_{x-y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j), \quad \text{con } |i - j| = k = 0, 1, \dots, N_g$$

1. Momento angular de segundo orden:

Esta medida da valores altos cuando en la matriz de co-ocurrencia tiene pocas entradas de gran magnitud, y es baja cuando todas las entradas son similares. Es una medida de la homogeneidad local.

$$f_1 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (p(i, j))^2$$

2. Contraste:

Es una medida de la variación local en una imagen. Tiene un valor alto cuando la región dentro de la escala de la ventana tiene un alto contraste.

$$f_2 = \sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (p(i, j)) \right\}$$

3. Correlación:

Un objeto tiene más alta correlación dentro de él que entre objetos adyacentes. Píxeles cercanos están más correlacionados entre sí que los píxeles más distantes.

$$f_3 = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (ij) p(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y}$$

Donde σ_x , σ_y , μ_x y μ_y son la media y el desvío estándar de p_x y p_y

4. Varianza:

$$f_4 = \sum_{i=1}^{N_g} \sum_j^{N_g} (i - \mu)^2 p(i, j)$$

5. Homogeneidad:

La homogeneidad es alta cuando la matriz de co-ocurrencia se concentra a lo largo de la diagonal. Esto ocurre cuando la imagen es localmente homogénea de acuerdo al tamaño de la ventana.

$$f_5 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{1}{1 + \frac{1}{(i-j)^2}} p(i, j)$$

6. Suma de la media:

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

7. Suma de la varianza:

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

8. Suma de la entropía:

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

9. Entropía:

$$f_9 = - \sum_{i=1}^{N_g} \sum_j^{N_g} p(i, j) \log(p(i, j))$$

10. Varianza diferencial:

$$f_{10} = varianza de p_{x-y}$$

11. Entropía diferencial:

$$f_{11} = - \sum_{i=0}^{N_g-1} p_{x-y}(i) \log(p_{x-y}(i))$$

12. Medidas de información de la correlación:

$$f_{12} = \frac{HXY}{\max\{HX, HY\}} - \frac{HXY1}{HY}$$

$$f_{13} = [1 - \exp(-2(HXY2 - HXY))]^{1/2}$$

$$HXY = - \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j) \log(p(i, j))$$

$$HXY1 = - \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j) \log(p_x(i) p_y(j))$$

$$HXY2 = - \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_x(i) p_y(j) \log(p_x(i) p_y(j))$$

13. Coeficiente de correlación máximo:

$$f_{14} = (\text{Segundo mayor valor propio de } Q)^{1/2}$$

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

2.11. Reconocimiento de patrones

El reconocimiento de patrones tiene como objetivo la clasificación de objetos en un número de clases dado. Un patrón es un punto en el espacio de representación, cuya dimensión viene determinada por el número de variables (características) consideradas.

Las propiedades que deben cumplir las características son:

- Capacidad discriminante: deben separar lo más nítidamente posibles las clases existentes.
- Fiabilidad: los objetos de una misma clase tendrán la menor dispersión posible.
- Incorrelación: las características no dependerán fuertemente de otras, ya que si fuera así, no aportarían información.

Si se considera que cada clase tiene asociado un agrupamiento bien diferenciado de las demás, el problema de la clasificación se reduce a buscar superficies que separen los distintos agrupamientos.

En esta sección se desarrollará el funcionamiento de diversos clasificadores utilizados. Se añade además un análisis que permite reducir la dimensionalidad del conjunto de datos (análisis de componentes principales).

2.11.1. Máquinas de vectores de soporte

Las máquinas de vectores soporte (SVM, del inglés Support Vector Machines) fueron pensadas para resolver problemas de clasificación binaria, pero actualmente se utilizan para resolver otros tipos de problemas como regresión, agrupamiento, multiclasicación.[\[26\]](#)

Dentro de la tarea de clasificación, las SVMs pertenecen a la categoría de los clasificadores lineales, puesto que inducen separadores lineales o hiperplanos, ya sea en el espacio original de los ejemplos de entrada, si éstos son separables o quasi-separables (ruido), o en un espacio transformado (espacio de características), si los ejemplos no son separables linealmente en el espacio original. La búsqueda del hiperplano de separación en estos espacios transformados, normalmente de muy alta dimensión, se hará de forma implícita utilizando las denominadas funciones kernel.

El sesgo inductivo asociado a las SVMs radica en la minimización del denominado riesgo estructural. La idea es seleccionar un hiperplano de separación que equidista de los ejemplos más cercanos de cada clase para conseguir un margen máximo a cada lado del hiperplano. Además, a la hora de definir el hiperplano, sólo se consideran los ejemplos de entrenamiento de cada clase que caen justo en la frontera de dichos márgenes. Estos ejemplos reciben el nombre de vectores soporte. Desde un punto de vista práctico, el hiperplano separador de margen máximo ha demostrado tener una buena capacidad de generalización, evitando en gran medida el problema del sobreajuste a los ejemplos de entrenamiento.

Desde un punto de vista algorítmico, el problema de optimización del margen geométrico representa un problema de optimización cuadrático con restricciones lineales que puede ser resuelto mediante técnicas estándar de programación cuadrática. La propiedad de convexidad exigida para su resolución garantizan una solución única.

Definición de SVM para clasificación binaria de ejemplos linealmente separables

Dado un conjunto separable de ejemplos $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$, donde $x_i \in \mathbb{R}^d$ e $y_i \in \{+1, -1\}$, se puede definir un hiperplano de separación lineal que es capaz de separar dicho conjunto sin error:

$$D(x) = (w_1x_1 + \dots + w_dx_d) + b = \langle w, x \rangle + b \quad (2.1)$$

donde w y b son coeficientes reales.

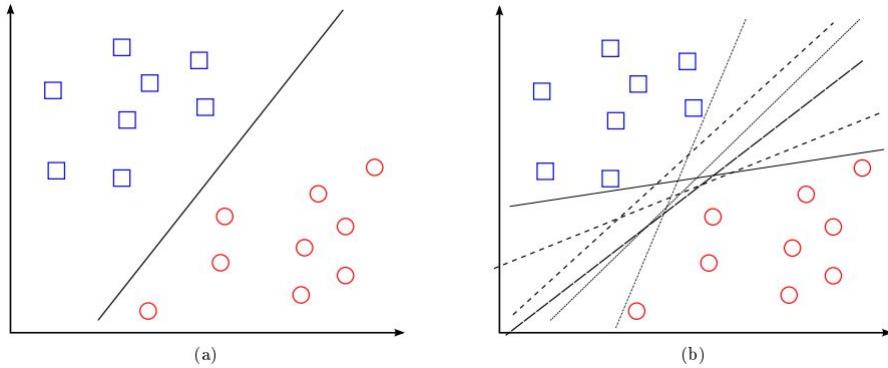
El hiperplano de separación cumplirá las siguientes restricciones para todo x_i del conjunto de ejemplos:

$$\langle w, x_i \rangle + b \geq 0 \quad \text{si } y_i = +1$$

$$\langle w, x_i \rangle + b \leq 0 \quad \text{si } y_i = -1, \quad i = 1, \dots, n \quad (2.2)$$

O también:

$$y_i(\langle w, x_i \rangle + b) \geq 0, \quad i = 1, \dots, n \quad (2.3)$$



De forma más compacta:

$$y_i D(x_i) \geq 0, \quad i = 1, \dots, n \quad (2.4)$$

Como se puede ver en la Figura 2.11.2, el hiperplano que permite separar los ejemplos no es único, es decir, existen infinitos hiperplanos separables, representados por todos aquellos hiperplanos que son capaces de cumplir las restricciones impuestas por cualquiera de las expresiones equivalentes recién mencionadas. Para establecer un criterio adicional que permita definir un hiperplano de separación óptimo, primero se define el concepto de margen de un hiperplano de separación, denotado por τ , como la mínima distancia entre el mismo y el ejemplo más cercano de cualquiera de las dos clases. Por lo tanto, un hiperplano de separación se denominará óptimo si su margen es de tamaño máximo.

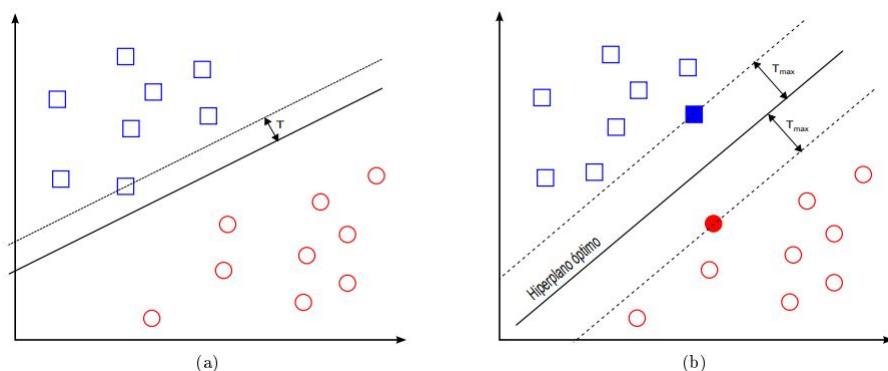


Figura 2.11.2: Márgenes de hiperplanos de separación. a) hiperplano de separación no-óptimo y su margen asociado (no máximo), b) hiperplano de separación óptimo y su margen asociado (máximo).

Una propiedad inmediata de esta definición es que el hiperplano óptimo equidista del ejemplo más cercano de cada clase. Por geometría, se sabe que la distancia entre un hiperplano de separación $D(x)$ y un ejemplo x' viene dada por:

$$\frac{|D(x')|}{\|w\|} \quad (2.5)$$

Siendo w el vector que, junto con el parámetro b , define el hiperplano $D(x)$ y que, además, tiene la propiedad de ser perpendicular al hiperplano considerado. Haciendo uso de la expresiones (2.4) y (2.5), todos los ejemplos de entrenamiento cumplirán que:

$$\frac{y_i |D(x')|}{\|w\|} \geq \tau, \quad i = 1, \dots, n \quad (2.6)$$

De la expresión anterior, se deduce que encontrar el hiperplano óptimo es equivalente a encontrar el valor de w que maximiza el margen. Sin embargo, existen infinitas soluciones que difieren solo en la escala de w . Así, por ejemplo, todas las funciones lineales $\lambda (< w, x > + b)$, con $\lambda \in \{R\}$, representan el mismo hiperplano. Para limitar, por tanto, el número de soluciones a una sola, y teniendo en cuenta que (2.6) se puede expresar también como:

$$y_i D(x_i) \geq \tau \|w\|, \quad i = 1, \dots, n \quad (2.7)$$

La escala del producto de τ y la norma de w se fija, de forma arbitraria, a la unidad, es decir:

$$\tau \|w\| = 1 \quad (2.8)$$

Llegando a la conclusión final de que aumentar el margen es equivalente a disminuir la norma de w , ya que la ecuación anterior se puede expresar como:

$$\tau = \frac{1}{\|w\|} \quad (2.9)$$

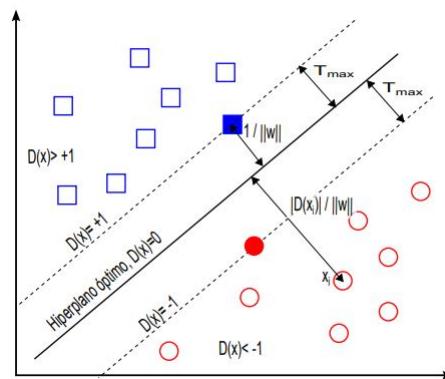


Figura 2.11.3: La distancia de cualquier ejemplo x_i al hiperplano de separación óptimo viene dada por $|D(x_i)|/\|w\|$. En particular, si dicho ejemplo pertenece al conjunto de vectores soporte (identificados por siluetas sólidas), la distancia a dicho hiperplano será siempre $1/\|w\|$. Además, los vectores soporte aplicados a la función de decisión siempre cumplen que $|D(x)| = 1$.

Un hiperplano de separación óptimo (Figura 2.11.3) será aquel que posee un margen máximo y, por tanto, un valor mínimo de $\|w\|$ y, además, está sujeto a la restricción dada por (2.7), junto con el criterio expresado por (2.8), es decir:

$$y_i |D(x)| \geq 1, \quad i = 1, \dots, n \quad (2.10)$$

O lo que es lo mismo:

$$y_i (\langle w, x_i \rangle + b) \geq 1, \quad i = 1, \dots, n \quad (2.11)$$

El concepto de margen máximo está relacionado directamente con la capacidad de generalización del hiperplano de separación, de tal forma que, a mayor margen, mayor distancia de separación existirá entre las dos clases. Los ejemplos que están situados a ambos lados del hiperplano óptimo y que definen el margen o, lo que es lo mismo, aquellos para los que la restricción (2.11) es una igualdad, reciben el nombre de vectores soporte. Puesto que estos ejemplos son los más cercanos al hiperplano de separación, serán los más difíciles de clasificar y, por tanto, deberían ser los únicos ejemplos a considerar a la hora de construir dicho hiperplano. De hecho, se demostrará más adelante, que el hiperplano de separación óptimo se define sólo a partir de estos vectores.

Este problema de optimización con restricciones corresponde a un problema de programación cuadrática y es abordable mediante la teoría de la optimización. Dicha teoría establece que un problema de optimización, denominado primal, tiene una forma dual si la función a optimizar y las restricciones son funciones estrictamente convexas. En estas circunstancias, resolver el problema dual permite obtener la solución del problema primal.

SVM para la clasificación binaria de ejemplos linealmente cuasi-separables

El problema planteado en la sección anterior tiene escaso interés práctico porque los problemas reales se caracterizan normalmente por poseer ejemplos ruidosos y no ser perfecta y linealmente separables. La estrategia para este tipo de problemas reales es relajar el grado de separabilidad del conjunto de ejemplos, permitiendo que haya errores de clasificación en algunos de los ejemplos del conjunto de entrenamiento. Sin embargo, sigue siendo un objetivo el encontrar un hiperplano óptimo para el resto de ejemplos que sí son separables.

Un ejemplo es no-separable si no cumple la condición (2.11). Aquí se pueden dar dos casos:

1. El ejemplo cae dentro del margen asociado a la clase correcta, de acuerdo a la frontera de decisión que define el hiperplano de separación. Para este caso la clasificación es correcta.
2. El ejemplo cae al otro lado de dicho hiperplano, por lo tanto la clasificación no es correcta.

La idea para abordar este nuevo problema es introducir un conjunto de variables reales positivas, denominadas variables de holgura, $\xi_i, \quad i = 1, \dots, n$, que permitirán cuantificar el número de

ejemplos no-separables que se está dispuesto a admitir, es decir:

$$y_i (\langle w, x_i \rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \dots, n \quad (2.12)$$

Por tanto, para un ejemplo (x_i, y_i) , su variable de holgura, ξ_i , representa la desviación del caso separable, medida desde el borde del margen que corresponde a la clase y_i , ver Figura 2.11.4.

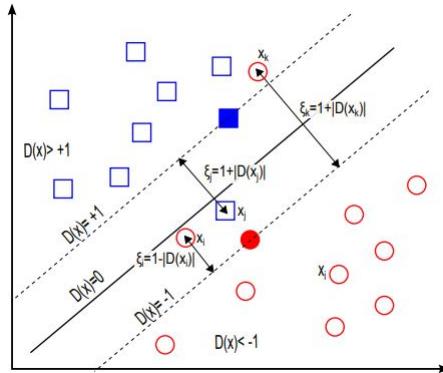


Figura 2.11.4: Obtención de hiperplano óptimo en ejemplos no separables.

Cuanto mayor sea el valor de la suma de todas las variables de holgura, mayor será el número de ejemplos no separables.

Relajadas las restricciones, ya no basta con plantear como único objetivo maximizar el margen, ya que podríamos lograrlo a costa de clasificar erróneamente muchos ejemplos. Por tanto, la función a optimizar debe incluir los errores de clasificación que está cometiendo el hiperplano de separación:

$$f(w, \xi) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i \quad (2.13)$$

Donde C es una constante, suficientemente grande, elegida por el usuario, que permite controlar en qué grado influye el término del coste de ejemplos no-separables en la minimización de la norma, es decir, permitirá regular el compromiso entre el grado de sobreajuste del clasificador final y la proporción del número de ejemplos no separables. Así, un valor de C muy grande permitiría valores de ξ_i muy pequeños. En el límite $C \rightarrow \infty$, estaríamos considerando el caso de ejemplos perfectamente separables $\xi_i \rightarrow 0$.

En consecuencia, el nuevo problema de optimización consistirá en encontrar el hiperplano, definido por w y b , que minimiza el funcional (2.13) y sujeto a las restricciones dadas por (2.12).

SVM para la clasificación de ejemplos linealmente no separables

Para este caso se utilizarán funciones base no lineales para definir espacios transformados de alta dimensionalidad y luego buscar hiperplanos de separación óptimos en dichos espacios transformados. A cada uno de estos espacios se le denomina espacio de características, para diferenciarlo del espacio de ejemplos de entrada (espacio- x).

Sea $\Phi : \mathbb{X} \rightarrow \mathbb{F}$ la función de transformación que hace corresponder cada vector de entrada x con un punto en el espacio de características \mathbb{F} , donde $\Phi = [\phi_1(x), \dots, \phi_m(x)]$ y $\exists \phi_i(x), i = 1, \dots, m$ tal que $\phi_i(x)$ es una función no lineal. La idea entonces es construir un hiperplano de separación lineal en este nuevo espacio. La frontera de decisión lineal obtenida en el espacio de características se transformará en una frontera de decisión no lineal en el espacio original de entradas, tal como podemos ver en la Figura 2.11.5.

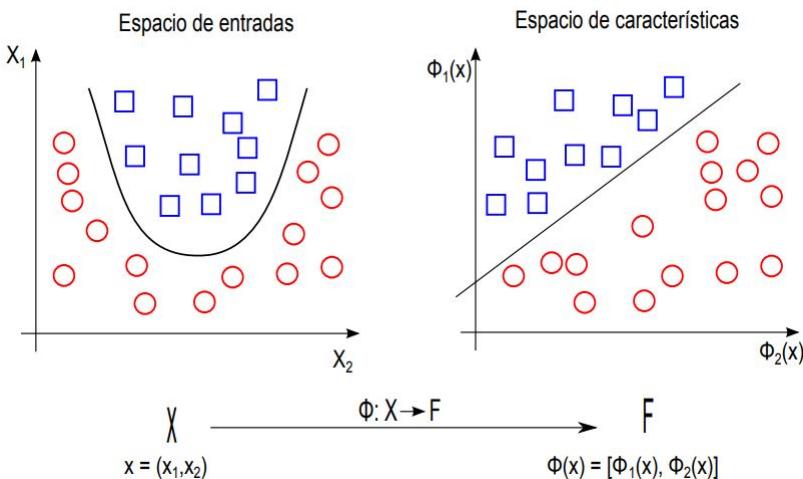


Figura 2.11.5: El problema de la búsqueda de una función de decisión no lineal en el espacio de entradas, se puede transformar en un nuevo problema consistente en la búsqueda de una función de decisión lineal (hiperplano) en un nuevo espacio transformado (espacio de características).

Operando según la teoría de la optimización es posible obtener la función de decisión en su forma dual:

$$D(x) = \sum_{i=1}^n \alpha_i^* y_i K(x, x_i) \quad (2.14)$$

Donde $K(x, x')$ se denomina función kernel.

Por definición, una función kernel es una función $\mathbb{K} : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ que asigna a cada par de elementos del espacio de entrada, \mathbb{X} , un valor real correspondiente al producto escalar de las imágenes de dichos elementos en un nuevo espacio \mathbb{F} (espacio de características).

Así, dado el conjunto de funciones base, $\Phi = \{\phi_1(x), \dots, \phi_m(x)\}$, el problema a resolver sigue siendo encontrar el valor de los parámetros $\alpha_i^*, i = 1, \dots, n$ que optimiza el problema dual.

Como consecuencia del Teorema de Aronszajn⁶, es posible afirmar que para resolver el problema dual, no sólo no se necesita conocer el conjunto de funciones base de transformación, sino que tampoco es necesario conocer las coordenadas de los ejemplos transformados en el espacio de características. Sólo se necesitará conocer la forma funcional del kernel correspondiente, aún cuando este

pudiera estar asociado a un conjunto infinito de funciones base.

Ejemplos de funciones kernel:

- Kernel lineal:

$$K(x, x') = \langle x, x' \rangle \quad (2.15)$$

- Kernel polinómico de grado p:

$$K_p(x, x') = [\gamma \langle x, x' \rangle + \tau]^p \quad (2.16)$$

- Kernel gaussiano:

$$K(x, x') = \exp(\gamma \|x - x'\|^2), \quad \gamma > 0 \quad (2.17)$$

- Kernel sigmoidal:

$$K(x, x') = \tanh(\gamma \langle x, x' \rangle + \tau) \quad (2.18)$$

A γ , τ y p se los denomina parámetros del kernel.

2.11.2. k-vecinos más próximos (k-NN)

En inglés, k-Nearest Neighbors, abreviado k-NN es un método de clasificación supervisada. Este es un método de clasificación no paramétrico, que estima el valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento x pertenezca a la clase C a partir de la información proporcionada por el conjunto de ejemplos.

La idea básica sobre la que se fundamenta este paradigma es que un nuevo caso se va a clasificar en la clase más frecuente a la que pertenecen sus K vecinos más cercanos.[\[27\]](#)

Siendo D un fichero de N casos, cada uno de los cuales está caracterizado por n variables predictoras, X_1, \dots, X_n y una variable a predecir, la clase C.

Los N casos se denotan por:

$(x_1, c_1), \dots, (x_n, c_n)$ donde:

$x_i = (x_{i,1}, \dots, x_{i,n})$ para todo $i=1, \dots, N$

$c_i \in c^1, \dots, c^m$ para todo $i=1, \dots, N$

c^1, \dots, c^m denotan los posibles valores de la variable clase C.

⁶Teorema de Aronszajn. Para cualquier función $K : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ que sea simétrica y semidefinida positiva, existe un espacio de Hilbert y una función $\Phi : \mathbb{X} \rightarrow \mathbb{F}$ tal que: $K(x, x') = \langle \Phi(x), \Phi(x') \rangle \quad \forall x, x' \in X$

| | | X_1 | ... | X_j | ... | X_n | C |
|-----------------------|----------|-------------|-----|-------------|-----|-------------|----------|
| (\mathbf{x}_1, c_1) | 1 | x_{11} | ... | x_{1j} | ... | x_{1n} | c_1 |
| | \vdots | \vdots | | \vdots | | \vdots | \vdots |
| (\mathbf{x}_i, c_i) | i | x_{i1} | ... | x_{ij} | ... | x_{in} | c_i |
| | \vdots | \vdots | | \vdots | | \vdots | \vdots |
| (\mathbf{x}_N, c_N) | N | x_{N1} | ... | x_{Nj} | ... | x_{Nn} | c_N |
| \mathbf{x} | $N + 1$ | $x_{N+1,1}$ | ... | $x_{N+1,j}$ | ... | $x_{N+1,n}$ | ? |

Figura 2.11.6: Ejemplo de notación para algoritmo k-NN.

El nuevo caso que se pretende clasificar se denota por $x = (x_1, \dots, x_n)$.

COMIENZO

Entrada: $D = \{(\mathbf{x}_1, c_1), \dots, (\mathbf{x}_N, c_N)\}$

$\mathbf{x} = (x_1, \dots, x_n)$ nuevo caso a clasificar

PARA todo objeto ya clasificado (\mathbf{x}_i, c_i)

calcular $d_i = d(\mathbf{x}_i, \mathbf{x})$

Ordenar $d_i (i = 1, \dots, N)$ en orden ascendente

Quedarnos con los K casos D_x^K ya clasificados más cercanos a \mathbf{x}

Asignar a \mathbf{x} la clase más frecuente en D_x^K

FIN

Figura 2.11.7: Pseudocódigo para el clasificador k-NN.

En la Figura 2.11.7 se presenta un pseudocódigo para el clasificador k-NN básico. Tal y como puede observarse en el mismo, se calculan las distancias de todos los casos ya clasificados al nuevo caso, x , que se pretende clasificar. Una vez seleccionados los K casos ya clasificados, D_x^K más cercanos al nuevo caso, x , a éste se le asignará la clase (valor de la variable C) más frecuente de entre los K objetos, D_x^K .

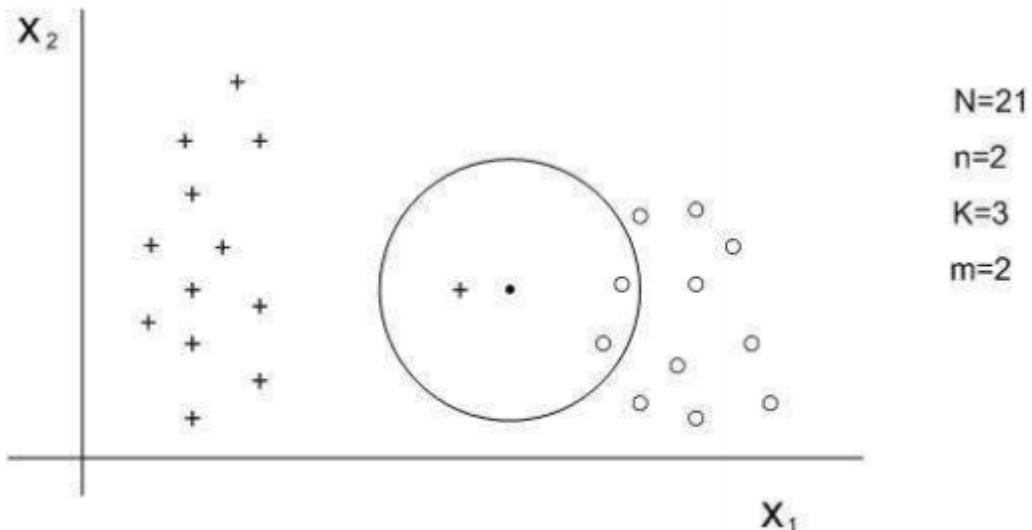


Figura 2.11.8: Ejemplo de aplicación del algoritmo k-NN básico.

En la Figura 2.11.8 hay veinticuatro (24) casos ya clasificados en dos posibles valores ($m = 2$). Las variables predictoras son X_1 y X_2 , y se ha seleccionado $K = 3$. De los 3 casos ya clasificados que se encuentran más cercanos al nuevo caso a clasificar, x (representado por ●), dos de ellos pertenecen a la clase ○, por tanto el clasificador 3-NN predice la clase ○ para el nuevo caso. Nótese que el caso más cercano a x pertenece a la clase +. Es decir, que si hubiésemos utilizado un clasificador 1-NN, x se hubiese asignado a +.

En caso de que se produzca un empate entre dos o más clases, conviene tener una regla heurística para su ruptura. Ejemplos de reglas heurísticas para la ruptura de empates pueden ser: seleccionar la clase que contenta al vecino más próximo, seleccionar la clase con distancia media menor, etc.

Otra cuestión importante es la determinación del valor de K . Se constata empíricamente que el porcentaje de casos bien clasificados es no monótono con respecto de K , siendo una buena elección valores de K comprendidos entre 3 y 7.

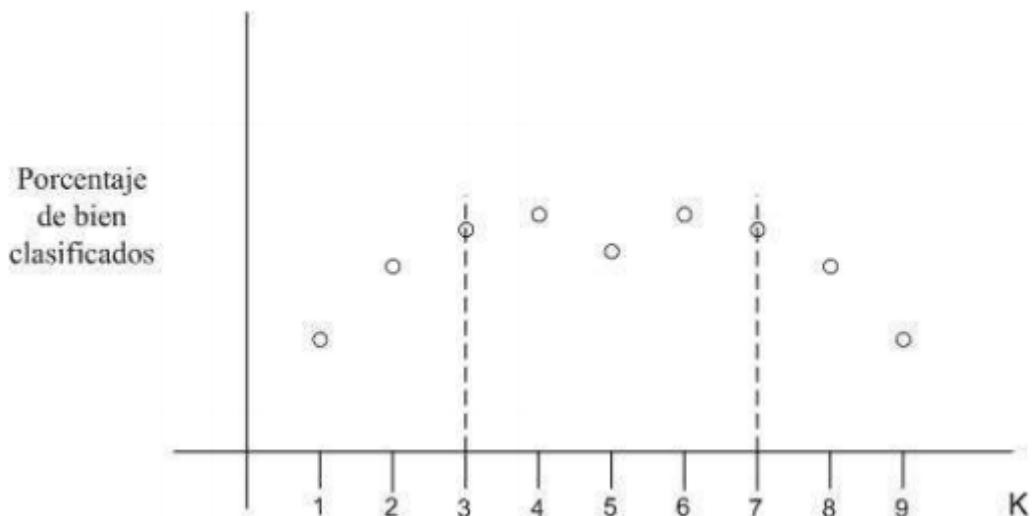


Figura 2.11.9: Ejemplo de la no monotonicidad del porcentaje de bien clasificados en función de K .

2.11.3. Regresión logística

La regresión logística es un tipo de análisis utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras.[28]

Si se supone que la variable dependiente Y representa la ocurrencia o no de un suceso (toma valor 1 si ocurre y 0 si no), interesa estudiar la relación entre una o más variables independientes o explicativas: X_1, X_2, \dots, X_p y la variable Y .

El modelo logístico establece la siguiente relación entre la probabilidad de que ocurra el suceso, dado que el individuo presenta los valores $X_1 = x_1, X_2 = x_2, \dots, X_p = x_p$:

$$Pr(Y = 1 | x_1, x_2, \dots, x_p) = \frac{1}{1 + exp(-\alpha - \beta_1 x_1 - \beta_2 x_2 - \dots - \beta_p x_p)}$$

Otra forma de presentar esta relación es:

$$\text{logit}(\Pr(Y = 1|x)) = \log\left(\frac{\Pr(Y=1|x)}{1-\Pr(Y=1|x)}\right) = -\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

Donde denotamos con $\Pr(Y = 1|x)$ la probabilidad condicional $\Pr(Y = 1|x_1, x_2, \dots, x_p)$

El procedimiento de estimación de estos parámetros α, β_i se basa en el método de máxima verosimilitud. Existen varios programas que realizan estas estimaciones mediante la obtención del máximo del logaritmo de la función de verosimilitud:

$$L(y, \beta) = \sum_{i=1}^n y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)$$

donde n es el número de observaciones y $p_i = \Pr(Y = y_i|x_i)$

Para obtener el estimador máximo verosímil hay que seguir los siguientes pasos:

- Derivar el logaritmo de la función de verosimilitud respecto del parámetro desconocido.
- Igualar a cero la primera derivada y obtener el valor del estimador en función de los elementos de la muestra.
- Obtener la segunda derivada del logaritmo de la función de verosimilitud respecto del parámetro. Para que sea un máximo esta segunda derivada tiene que ser negativa.

Una vez obtenidos los parámetros se calcula la probabilidad Pr de que el sujeto este encuadrado en esa categoría es mayor que 0,5 se le asigna, si es menor se le asignara la otra categoría.

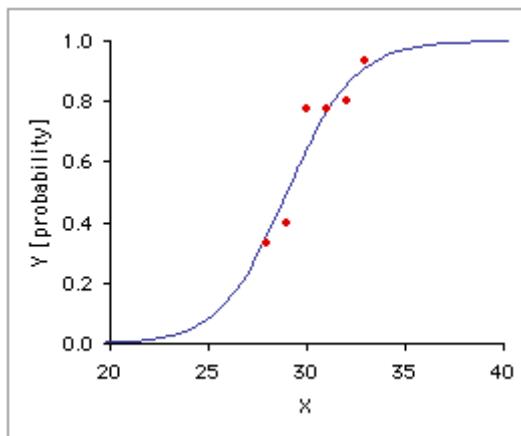


Figura 2.11.10: Ejemplo de la no monotonicidad del porcentaje de bien clasificados en función de K.

Las variables explicativas de tipo nominal con más de dos categorías deben ser incluidas en el modelo definiendo variables dummy. Si el sistema M categorías deben crearse entonces ($M1$) variables dicotómicas que son las llamadas variables dummy asociadas a la variable nominal. Las ($M1$) variables dicotómicas se denotan por $(Z_1, Z_2, \dots, Z_{M-1})$. A cada categoría o clase de la variable nominal le corresponde un conjunto de valores de los Z_1 con el cual se identifica dicha clase.

La manera más usual de definir estas (M_1) variables es la siguiente: si el sujeto pertenece a la primera categoría, entonces las (M_1) variables dummy valen ($Z_1 = 0 = Z_2, \dots = Z_{M-1} = 0$) ; si el sujeto se halla en la segunda categoría, ($Z_1 = 1, Z_2, \dots = Z_{M-1}$) ; si el sujeto se halla en la tercera categoría, ($Z_1 = 0, Z_2 = 1, \dots Z_{M-1} = 0$) ; y así sucesivamente hasta llegar a la última categoría, para la cual $Z_{M-1} = 1$ y las restantes valen 0.

2.11.4. Random forest

El bosque aleatorio crea múltiples árboles de decisión y los combina para obtener una predicción más precisa y estable. Un árbol de decisión es un modelo de predicción. Dado un conjunto de datos se fabrican diagramas de construcciones lógicas, muy similares a los sistemas basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema. Cuando llega el momento de predecir, el bosque aleatorio toma un promedio de todas las estimaciones individuales del árbol de decisión. Esto aumenta la diversidad en el bosque, lo que lleva a predicciones generales más sólidas, debido a que si hacemos una equivalencia con la realidad podemos pensar que si cientos de seres humanos hacen estimaciones para un problema en particular, al agrupar las predicciones, se podrá incorporar mucho más conocimiento que de cualquier individuo. [29]

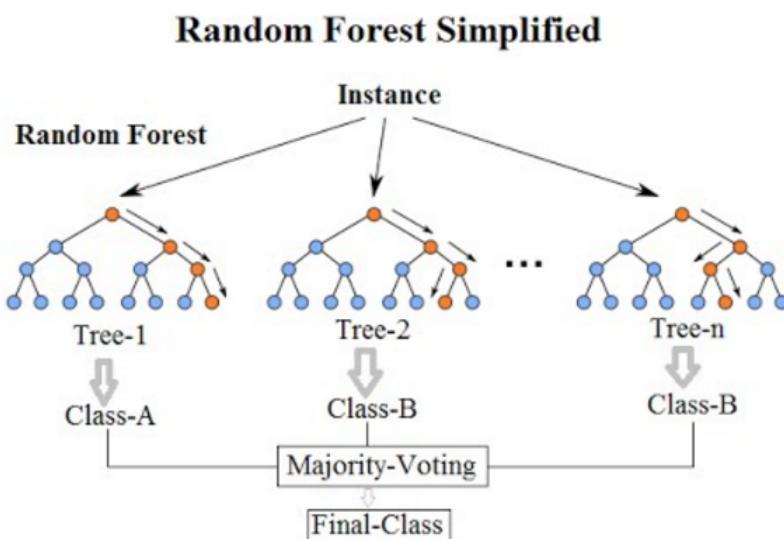


Figura 2.11.11: Algoritmo Random Forest. [29]

2.11.5. Análisis de componentes principales (ACP)

Un problema central en el análisis de datos multivariantes es la reducción de la dimensionalidad: si es posible describir con precisión los valores de p variables por un pequeño subconjunto $r < p$ de ellas, se habrá reducido la dimensión del problema a costa de una pequeña pérdida de información.[30]

El análisis de componentes principales tiene este objetivo: dadas n observaciones de p variables, se

analiza si es posible representar adecuadamente esta información con un número menor de variables incorreladas construidas como combinaciones lineales de las originales.

Supongamos que se dispone de los valores de p-variables en n elementos de una población dispuestos en una matriz X de dimensiones n x p, donde las columnas contienen las variables y las filas los elementos. El paso previo corresponde restarle a cada variable su media, de manera que las variables de la matriz X tienen media cero y su matriz de covarianzas vendrá dada por $\frac{1}{n}X'X$.

Se desea encontrar un subespacio de dimensión menor que p tal que al proyectar sobre él los puntos conserven su estructura con la menor distorsión posible. En el caso de p=2:

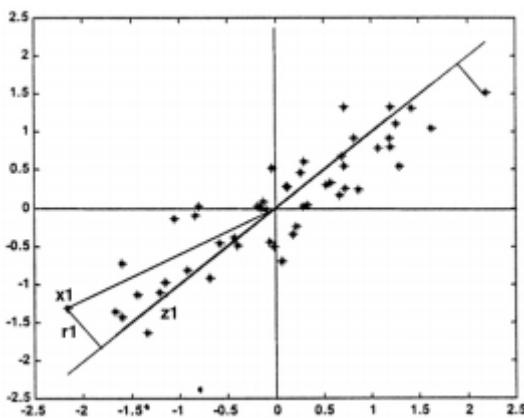


Figura 2.11.12: Ejemplo de la recta que minimiza las distancias ortogonales de los puntos a ella.

La Figura 2.11.12 indica el diagrama de dispersión y una recta que, intuitivamente, proporciona un buen resumen de los datos, ya que la recta pasa cerca de todos los puntos y las distancias entre ellos se mantienen aproximadamente en su proyección sobre la recta.

La condición de que la recta pase cerca de la mayoría de los puntos puede concretarse exigiendo que las distancias entre los puntos originales y sus proyecciones sobre la recta sean lo más pequeñas posibles. En consecuencia, si consideramos un punto x_i y una dirección $a_1 = (a_{11}, \dots, a_{1p})'$, definida por un vector a_1 de norma unidad, la proyección del punto x_i sobre esta dirección es el escalar:

$$z_i = a_{11}x_{i1} + \dots + a_{1p}x_{ip} = a_1'x_i$$

Y el vector que representa esta proyección será $z_i a_1$. Llamando r_i a la distancia entre el punto x_i , y su proyección sobre la dirección a_1 , este criterio implica:

$$\text{minimizar} \quad \sum_{i=1}^n r_i^2 = \sum_{i=1}^n |x_i - z_i a_1|^2$$

Donde $|u|$ es la norma euclídea o módulo del vector u.

La Figura 2.11.12 muestra que al proyectar cada punto sobre la recta se forma un triángulo rectángulo donde la hipotenusa es la distancia del punto al origen, $(x_i'x_i)^{\frac{1}{2}}$, y los catetos la proyección del punto sobre la recta z_i y la distancia entre el punto y su proyección (r_i). Por el teorema de Pitágoras, podemos escribir:

$$x_i' x_i = z_i^2 + r_i^2$$

Sumando esta expresión para todos los puntos, se obtiene:

$$\sum_{i=1}^n x_i' x_i = \sum_{i=1}^n z_i^2 + \sum_{i=1}^n r_i^2$$

Como el primer miembro es constante, minimizar $\sum_{i=1}^n r_i^2$, la suma de las distancias a la recta de todos los puntos, es equivalente a maximizar $\sum_{i=1}^n z_i^2$, la suma al cuadrado de los valores de las proyecciones. Como las proyecciones z_i son variables de media cero, maximizar la suma de sus cuadrados equivale a maximizar su varianza, y obtenemos el criterio de encontrar la dirección de proyección que maximice la varianza de los datos proyectados.

Cálculo del primer componente

El primer componente principal se define como la combinación lineal de las variables originales que tiene varianza máxima. Los valores en este primer componente de los n individuos se representarán por un vector z_1 , dado por:

$$z_1 = X a_1$$

Como las variables originales tienen media cero también z_1 tendrá media nula. Su varianza será:

$$\frac{1}{n} z_1' z_1 = \frac{1}{n} a_1' X' X a_1 = a_1' S a_1$$

Donde S es la matriz de varianzas y covarianzas de las observaciones. Es obvio que podemos maximizar la varianza sin límite aumentando el módulo del vector a_1 . Para que la maximización tenga solución debemos imponer una restricción al módulo del vector a_1 , y, sin pérdida de generalidad, impondremos que $a_1' a_1 = 1$

Introduciremos esta restricción mediante el multiplicador de Lagrange:

$$M = a_1' S a_1 - \lambda (a_1' a_1 - 1)$$

y maximizaremos esta expresión de la forma habitual derivando respecto a los componentes de a_1 e igualando a cero. Entonces:

$$\frac{\partial M}{\partial a_1} = 2S a_1 - 2\lambda a_1 = 0$$

cuya solución es:

$$S a_1 = \lambda a_1$$

Lo que implica que a_1 es un vector propio de la matriz S , y λ su correspondiente valor propio. Para determinar qué valor propio de S es la solución, multiplicando por la izquierda por a_1' esta ecuación:

$$a'_1 Sa_1 = \lambda a'_1 a_1 = \lambda$$

y concluimos, que λ es la varianza de z_1 . Como ésta es la cantidad que queremos maximizar, λ será el mayor valor propio de la matriz S. Su vector asociado, a_1 , define los coeficientes de cada variable en el primer componente principal.

Cálculo del segundo componente

Vamos a obtener el mejor plano de proyección de las variables X. Lo calcularemos estableciendo como función objetivo que la suma de las varianzas de $z_1 = Xa_1$ y $z_2 = Xa_2$ sea máxima, donde a_1 y a_2 son los vectores que definen el plano. La función objetivo será:

$$\phi = a'_1 Sa_1 + a'_2 Sa_2 - \lambda_1(a'_1 a_1 - 1) - \lambda_2(a'_2 a_2 - 1)$$

Que incorpora las restricciones de que las direcciones deben de tener módulo unitario $a'_i a_i = 1$, $i=1,2$. Derivando e igualando a cero:

$$\frac{\partial \phi}{\partial a_1} = 2Sa_1 - 2\lambda_1 a_1 = 0$$

$$\frac{\partial \phi}{\partial a_2} = 2Sa_2 - 2\lambda_2 a_2 = 0$$

La solución de este sistema es:

$$Sa_1 = \lambda_1 a_1$$

$$Sa_2 = \lambda_2 a_2$$

Que indica que a_1 y a_2 deben ser vectores propios de S. Tomando los vectores propios de norma uno se obtiene que, en el máximo, la función objetivo es:

$$\phi = \lambda_1 + \lambda_2$$

Es claro que λ_1 y λ_2 deben ser los dos autovalores mayores de la matriz S y a_1 y a_2 sus correspondientes autovectores. Observemos que la covarianza entre z_1 y z_2 , dada por $a'_1 S a_2$ es cero ya que a $a'_1 a_2 = 0$, y las variables z_1 y z_2 estarán incorreladas.

2.12. Estado del arte

En esta sección, se expondrán algunos trabajos previamente realizados:

2.12.1. Cuantificación de la densidad mamaria mediante Entropía de Permutaciones en mamografías [31]

- Materiales: Se obtuvieron 168 mamografías del banco de datos “Digital Database for Screening Mammography” (DDSM) para casos normales (sin tumores, benignos ni malignos). Procesamiento realizado en Matlab.
- Se seleccionó de cada mamografía la región de interés de estudio (tejido glandular). Se obtuvieron imágenes rectangulares de tamaños variados de aproximadamente 70x100 píxeles. Donde la región de interés está definida por los médicos expertos.

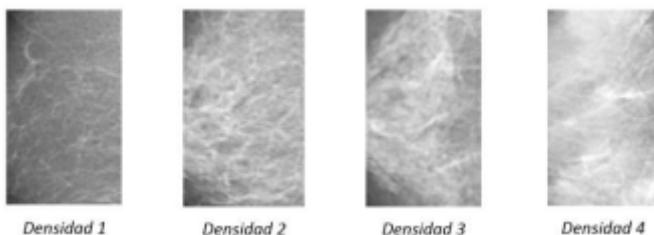


Figura 2.12.1: Tejido glandular de 4 mamografías que representan las distintas densidades. Imágenes reales procesadas por los algoritmos.[31]

- Metodología: Considerando la EP y sus variantes, se proponen 15 algoritmos para calcular descriptores de la rugosidad de una imagen en tonos de gris. Clasificación: Red neuronal multicapa.
- Resultados: 70 % de los datos para entrenamiento (118 imágenes), 15 % para test (25 imágenes) y 15 % (25 imágenes) para validación.

TABLA VI
MATRIZ CONFUSIÓN DE TEST

| MAMOGRAFÍA S... | Clasificada como D1 | Clasificada como D2 | Clasificada como D3 | Clasificada como D4 |
|---------------------------|---------------------|---------------------|---------------------|---------------------|
| <i>Etiquetada como D1</i> | 7 87,50 % | 0 0,00 % | 0 0,00 % | 0 0,00 % |
| <i>Etiquetada como D2</i> | 0 0,00 % | 3 100 % | 3 60,00 % | 0 0,00 % |
| <i>Etiquetada como D3</i> | 0 0,00 % | 0 0,00 % | 2 40,00 % | 1 11,11% |
| <i>Etiquetada como D4</i> | 1 12,50 % | 0 0,00 % | 0 0,00 % | 8 88,89 % |

Figura 2.12.2: Matriz de confusión de test.[31]

Validación cruzada de tipo Hold-out: error de test 24,42 %.

2.12.2. Breast Density Characterization using Texton Distributions[32]

- Materiales: El algoritmo fue evaluado en un set de 100 mamografías (Oxford Database), 25 por cada clase de BIRADS. Para determinar la clase a la que corresponde cada imagen hubo un acuerdo independiente en la clasificación de densidad por tres expertos radiólogos.
- Metodología: Se utilizó un descriptor de texturas, “texton spatial dependence matrix (TSDM)”.
- Resultados: Se utilizó 10 mamografías por clase para el entrenamiento y 15 para prueba.

| Accuracy% | BI-RADS I | BI-RADS II | BI-RADS III | BI-RADS IV |
|-------------------|-----------|------------|-------------|------------|
| 4 Density Classes | 86% | 93% | 80% | 66% |

Figura 2.12.3: Classification accuracy results. [32]

2.12.3. Automated analysis of mammographic densities [33]

- Materiales: 30 casos (estudios mamográficos digitales de alta resolución), se utilizaron las proyecciones derecha e izquierda craneal-caudal, para un total de 60 imágenes. Las mujeres eran todas pre menopáusicas y tenían una edad comprendida entre los 36 y los 49 años, con una edad media de 44 años. Los casos fueron revisados por dos radiólogos.
- Métodos: Análisis fractal y análisis de histograma. Se utilizó un clasificador bayesiano.
- Resultados:

| Comparison | Intra-class correlation |
|-----------------------------|-------------------------|
| Regional skewness versus R1 | 0.73 |
| Fractal versus R1 | 0.74 |
| Skewness/fractal versus R1 | 0.84 |

Figura 2.12.4: Coeficientes de correlación intraclase, que resumen la performance de la clasificación automática versus la del radiólogo (R1). [33]

2.12.4. Análisis de la densidad de mama asistido por ordenador [34]

- Materiales: Base DDSM y MIAS.
- Métodos: Diversas técnicas de análisis de histograma. Clasificador: K-vecinos.
- Resultados:

| Método | Porcentaje de error alcanzado |
|---|-------------------------------|
| Histograma global normalizado con reducción | 31.37% |
| Histograma global normalizado con PCA | 31.37% |
| Multiresolution Histogram | 35.09% |
| Multiresolution Histogram con PCA | 30.34% |
| Histograma por Regiones | 32.92% |
| Histograma por Regiones con PCA | 30.12 % |
| Combinación Multiresolution y Por Regiones | 36.65% |
| Selección de características | 23.91% |
| Propuesto (Histogramas Locales) | 21.74% |

Figura 2.12.5: Comparación métodos estudiados. [34]

2.12.5. Automated classification of parenchymal patterns in mammograms [35]

- Materiales: 615 seleccionadas del programa de detección de cáncer de mama en Nijmegen.
- Metodología: Análisis de histograma local. Clasificador: K-vecinos.
- Resultados: Todos los resultados de la clasificación se calcularon utilizando el método de dejar uno fuera, en el que cada muestra se clasifica por un clasificador que se entrena en todas las demás muestras. Los resultados se evalúan utilizando índices de error de clasificación y matrices de confusión. Para distinguir los errores de clasificación menores y mayores, se definieron dos tasas de error e_1 y e_2 . La tasa de error menor e_1 denota la fracción de las mamografías que se clasifican de manera diferente de la verdad básica por un solo grado de densidad. La mayor fracción de error e_2 está determinada por el número de clasificaciones que se desvían en más de un grado.

| Feature combination | e_1 | e_2 |
|---|-------|-------|
| S_1, S_2, S_3 | 0.41 | 0.062 |
| $S_1, S_2, S_3, Pe_1, Pe_2$ | 0.36 | 0.026 |
| $\sigma_1, \sigma_2, \sigma_3$ | 0.45 | 0.127 |
| $S_1, S_2, S_3, \sigma_1, \sigma_2, \sigma_3$ | 0.36 | 0.037 |
| $S_1, S_2, S_3, \sigma_1, \sigma_2, \sigma_3, Pe_1, Pe_2$ | 0.33 | 0.023 |
| $\bar{S}, \bar{\sigma}, Pe_1, Pe_2$ | 0.37 | 0.018 |

Figura 2.12.6: Las fracciones de error menores y mayores e_1 y e_2 de las clasificaciones se obtuvieron mediante diferentes combinaciones de características. [35]

2.12.6. Texture Descriptors applied to Digital Mammography [36]

- Materiales: Base de datos MIAS.
- Metodología: Se realizó una comparación entre tres descriptores de características diferentes: GLCM, máscaras de Laws y LBP. Clasificadores utilizados: k-vecinos (KNN), discriminante

de Fisher, clasificador (Fisher), análisis discriminante lineal (LDC) y algoritmo de máquina de vectores de soporte (SVM).

- Resultados:

| | GLCM | | Feature selection | |
|---------------|-------------|------------------|--------------------------|------------------|
| | Mean | Deviation | Mean | Deviation |
| KNN | 0.580 | 0.019 | 0.679 | 0.019 |
| Fisher | 0.634 | 0.026 | 0.671 | 0.014 |
| LDC | 0.658 | 0.023 | 0.714 | 0.018 |
| SVM | 0.669 | 0.008 | 0.751 | 0.009 |

| | LAWS | | Feature selection | |
|---------------|-------------|------------------|--------------------------|------------------|
| | Mean | Deviation | Mean | Deviation |
| KNN | 0.550 | 0.019 | 0.673 | 0.019 |
| Fisher | 0.456 | 0.022 | 0.594 | 0.015 |
| LDC | 0.583 | 0.018 | 0.694 | 0.017 |
| SVM | 0.628 | 0.009 | 0.721 | 0.008 |

| | LBP | | Feature selection | |
|---------------|-------------|------------------|--------------------------|------------------|
| | Mean | Deviation | Mean | Deviation |
| KNN | 0.626 | 0.010 | 0.717 | 0.011 |
| Fisher | 0.667 | 0.011 | 0.708 | 0.014 |
| LDC | 0.653 | 0.010 | 0.749 | 0.010 |
| SVM | 0.684 | 0.008 | 0.790 | 0.005 |

Figura 2.12.7: Porcentaje de clasificación correcta con y sin selección de características. [36]

Capítulo 3

Materiales y métodos

En este capítulo presentaremos la metodología utilizada en cada etapa del proyecto. Comentaremos todo el desarrollo llevado a cabo, junto con las soluciones propuestas para resolver inconvenientes, llegando así a cumplir con los objetivos planteados verificando la veracidad o no de la hipótesis.

3.1. Materiales

Para este trabajo se utilizó:

- Mamografías digitales de alta resolución, cortesía de la Fundación Carlos Oulton.
 - Colección de seis mil seiscientas trece (6613) imágenes debidamente anonimizadas, correspondientes a mil quinientos setenta y seis (1576) estudios diferentes.
 - Imágenes adquiridas con tres diferentes mamógrafos, todos ellos Hologic Selenia Dimensions.
 - El rango etario de las pacientes va de los 28 a los 91 años.
 - La cantidad de imágenes por clase (acr) son:
 - **a:** cuatrocientas setenta y tres (473).
 - **b:** dos mil setecientas tres (2703).
 - **c:** dos mil setecientas noventa y siete (2797).
 - **d:** cuatrocientas veintitres (423).
 - Imágenes con implantes mamarios: seiscientas ochenta y tres (683).
- Computadora portátil Dell Inspiron M531R-5535, con sistema operativo Windows 8.1 de 64-bit, con procesador AMD A-10-5745M a 2.10 GHz y 8Gb de RAM.
- Computadora portátil Dell Inspiron 3579, con sistema operativo Windows 10 de 64-bit, procesador Intel Core i7-8750H, 32Gb de RAM y placa de video dedicada Nvidia Geforce GTX 1050Ti.

- Como lenguaje base de programación se utilizó Python (3.6.5) desde la distribución de Anaconda 5.2.0 64-bit.
- Como visualizador de imágenes se utilizó Radiant DICOM Viewer 64-bit.
- Matlab R2013b.
- Base de datos DDSM (formato DICOM).

3.2. Etapas fundamentales

Como modo de organizar la información y el desarrollo del proyecto, dedicaremos esta sección a la presentación del diagrama de flujo representando el hilo conductor que engloba las distintas etapas en las cuales dividimos el análisis.



3.3. Comienzos

Al describir el marco teórico y plantear nuestros objetivos, podemos notar que nos introducimos en dos grandes áreas de la Ingeniería Biomédica. Siendo una de ellas el Procesamiento Digital de Imágenes.

Durante el transcurso de la carrera hemos aprendido a tratar imágenes digitales en el entorno de Matlab. Por este motivo nuestros primeros análisis y desarrollos han sido implementados allí.

Daremos una breve descripción de los objetivos alcanzados e inconvenientes encontrados, que luego actuaron como motor en la decisión de utilizar otros materiales (software e imágenes).

Al comienzo utilizamos como fuente la base de datos DDSM (Digital Database for Screening Mammography). No daremos muchos detalles aquí (nos extenderemos en el apéndice correspondiente), pero si haremos hincapié en un punto importante. Las imágenes mamográficas que constituyen esta base son mamografías analógicas digitalizadas. En éstas el ruido que se puede observar es mucho mayor que el de una mamografía digital y necesitamos eliminarlo antes de la etapa de segmentación, indicando un aumento en el costo computacional y una mayor sensibilidad a errores, debido a que cada una de las imágenes necesita un tratamiento determinado.

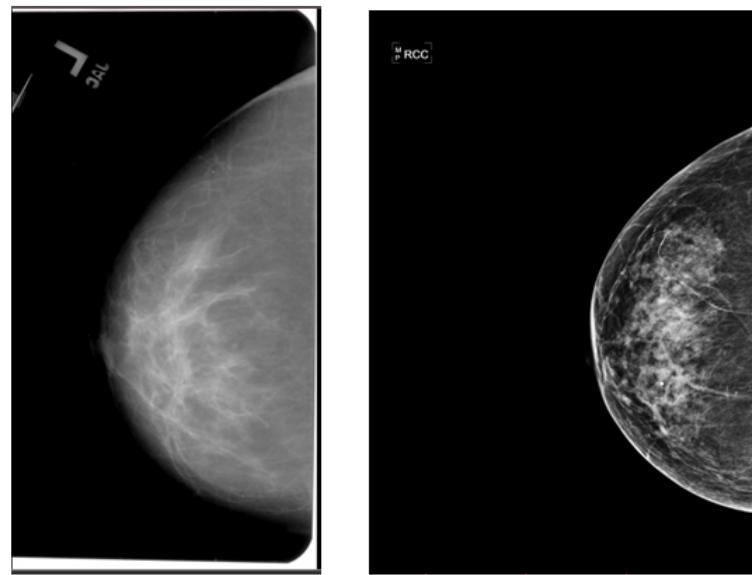


Figura 3.3.1: Imagen mamográfica analógica digitalizada (izquierda) Vs. Imagen digital(derecha).

En este caso agregaremos al diagrama de flujo presentado en la sección anterior una etapa previa a la segmentación necesaria para eliminar el ruido (sombras en la región de la mama debida al escaneado de la película, bordes blancos debido a su posicionamiento y utilización de filtros para mejorar la calidad de la imagen). Los resultados se podrán ver en la siguiente imagen.

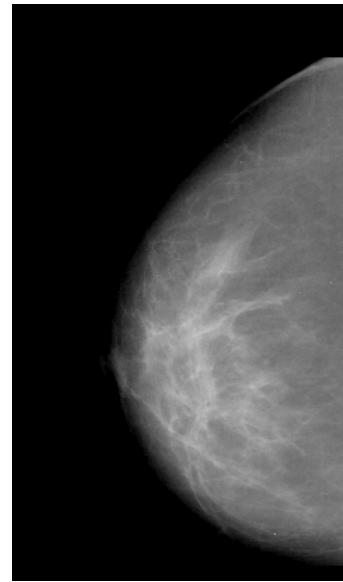


Figura 3.3.2: Imagen mamográfica analógica digitalizada luego de eliminar el ruido.

Este procedimiento se realizó automáticamente ya que la gran cantidad de condicionales incluidos dentro del código permitió que cada imagen tome el camino correcto para su tratamiento.

Luego de finalizar esa etapa nos adentramos en la de segmentación para obtener la región de interés (en nuestro caso la región compuesta por tejido mamario), por lo que el primer punto surge

de eliminar el músculo pectoral en las imágenes tomadas con una orientación oblicua medio-lateral (MLO).

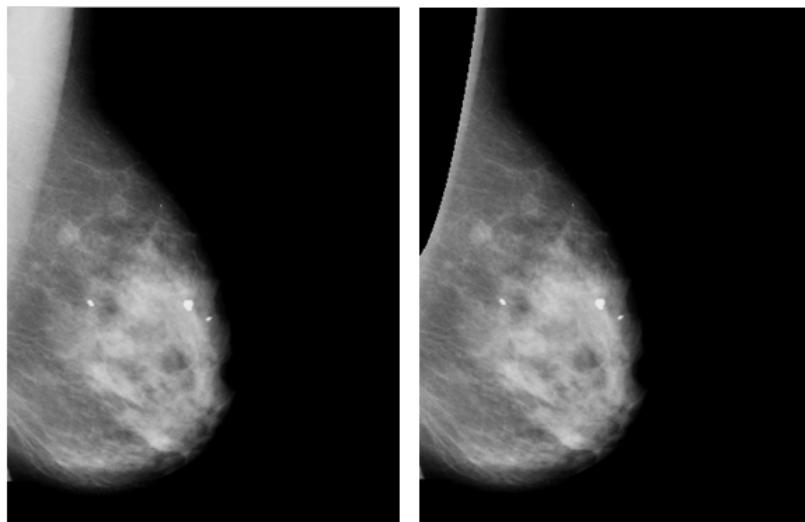


Figura 3.3.3: Extracción del músculo pectoral

El mecanismo automático de extraer el músculo nos ha traído varios problemas a lo largo del desarrollo del proyecto (serán expuestos más adelante). En esta sección solo diremos que si bien el resultado mostrado en la figura fue óptimo, ha fallado en gran cantidad de imágenes. Este procedimiento se llevó a cabo con técnicas de umbralización y ajuste de curva (polinomio de segundo grado). Posteriormente veremos como las combinamos con otros métodos para mejores resultados.

Si bien Matlab es un entorno de computación numérica fuertemente desarrollado, que junto con su caja de herramientas (toolbox) es muy utilizado para el procesamiento de imágenes, hemos decidido continuar utilizando como lenguaje de programación Python. La razón principal se debe a que éste posee una licencia de código abierto.

Otro cambio muy importante es que dejamos de utilizar las imágenes de la base DDSM, y comenzamos nuevamente el análisis con mamografías digitales extraídas del Instituto Oulton. La diferencia en calidad de imagen es la principal ventaja que obtuvimos al hacer este cambio, aunque no es la única, si no que también pudimos extraer características significativas de la información DICOM, tales como la edad de la paciente, la dosis recibida, el tiempo de exposición, los parámetros con los cuales fue ajustado el equipo para ese estudio (KV, mA), entre otros. Estas fueron utilizadas en la etapa de clasificación, dado que estos datos pueden ayudar en la determinación de la densidad radiológica.

3.4. Segmentación

Para analizar la densidad radiológica de la mama necesitamos quedarnos solamente con el tejido graso y glandular, a lo que llamaremos región de interés (ROI). Por lo tanto el objetivo de esta etapa es eliminar lo que no forma parte la misma.

Esta etapa fue dividida a su vez en cuatro subetapas:

- Eliminación de etiqueta.
- Sustracción de implante mamario.
- Extracción del músculo pectoral.
- Sustracción de piel.

3.4.1. Eliminación de etiqueta

Para eliminar la etiqueta superior que poseen las imágenes hemos aplicado respectivamente:

- Binarizado: utilizando un valor umbral 0.
- Etiquetado: quedándonos con la región pegada al borde.
- Multiplicación: imagen original con imagen etiquetada.

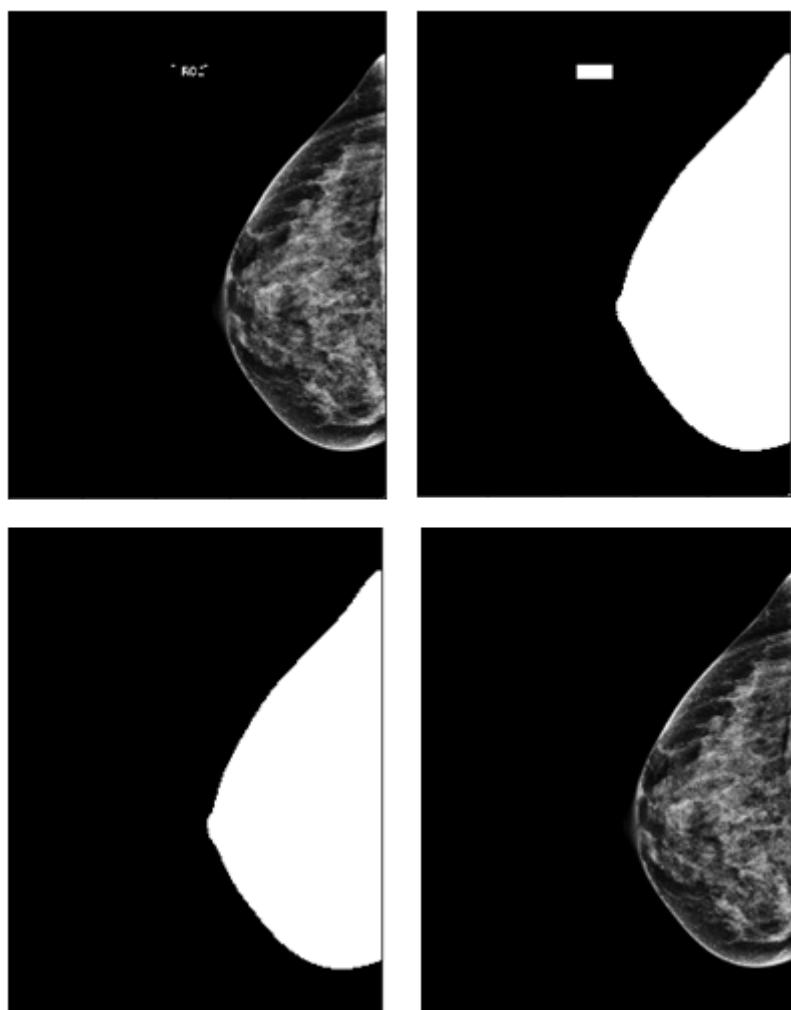


Figura 3.4.1: Eliminación de etiqueta

3.4.2. Sustracción de implante mamario

Dentro de la información DICOM, cada imagen posee un ítem donde se verifica la presencia o no de implante mamario (Breast Implant Present= “YES or NO”). Esto fue utilizado para responder al condicional de entrada del algoritmo.

Las imágenes con prótesis siguen el siguiente tratamiento:

- Crecimiento de Regiones: el criterio con el cual han sido elegidas las coordenadas de las semillas fue definido experimentalmente por medio de una serie de pruebas. Estas se posicionan alrededor de la coordenada $(\frac{3}{4} \text{ (filamax)}, \frac{1}{4} \text{ (columnamax)})$ para imágenes con Laterality = “L”, y $(\frac{3}{4} \text{ (filamax)}, \frac{3}{4} \text{ (columnamax-columnamin)+columnamin})$ para Laterality = “R”. Siendo filamax la fila máxima en la cual hay valores de pixeles mayores a cero, lo mismo para columnamax y columnamin.
- Negativo: se obtiene el negativo de la región resultante del método anterior.
- Dilatación: aplicada para llenar puntos negros.
- Multiplicación: imagen original con imagen negativa dilatada.

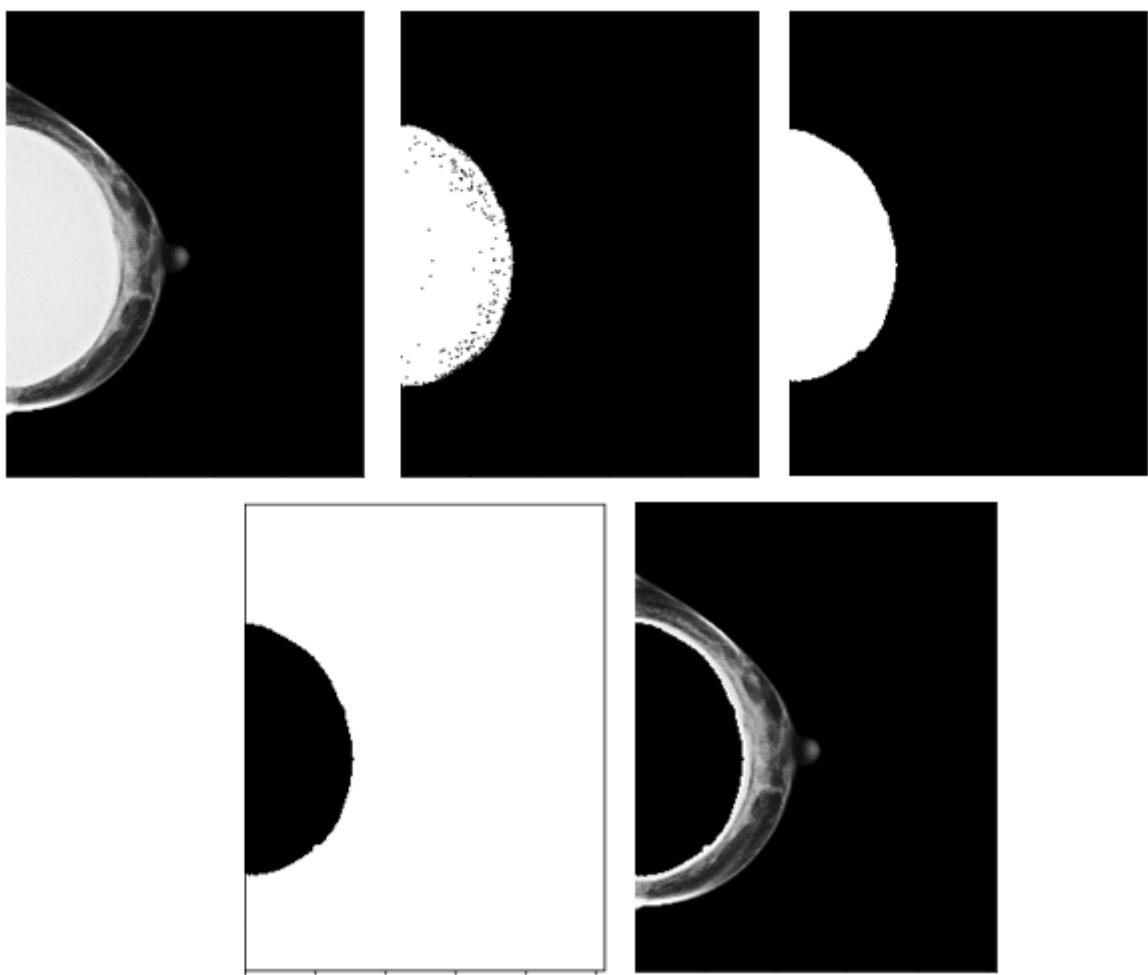


Figura 3.4.2: Sustracción de implante mamario

Se puede observar que en la imagen resultante el implante no ha sido del todo extraído. En la sección de extracción de piel se completará la tarea.

3.4.3. Extracción del músculo pectoral

Como se ha dicho anteriormente las imágenes que poseen músculo son las que fueron obtenidas en la orientación oblicuo mediolateral (MLO).

Esta etapa ha resultado la mas compleja debido a que partiendo de la hipótesis de ser el músculo la región de mayor intensidad (mas blanca) de la imagen, se propone como solución la aplicación de técnicas de umbralizado y ajuste de curva (tal como se ha desarrollado en Matlab). Si bien en la mayoría de los casos esto se cumple, existen excepciones en donde esta región es la menos intensa. A su vez en los casos positivos no siempre la aplicación de esos dos métodos devuelven un resultado óptimo. Esto será demostrado junto con la visualización de imágenes.

La solución propuesta, es un algoritmo automático constituido por una serie de métodos combinados y adaptados (mediante condiciones) a la variabilidad de imágenes que se pueden encontrar. Ellos son:

- Operador gradiente: utilizado como una forma de detectar bordes, entre ellos el del músculo.
- Suma: se realiza una suma ponderada para aumentar el contraste en los bordes.
- Umbralización de Ridler Calvard (modificada): se aplica este método para poder ubicar los puntos generadores en el crecimiento de regiones.
- Crecimiento de regiones : una vez ubicadas las semillas se realiza un crecimiento de región para luego detectar los puntos con los que se ajustará la curva.
- Ajuste de curva: esta técnica será explicada con mayor detalle ya que es la que mayor cantidad de adaptaciones posee en el algoritmo propuesto.
- Contorno: utilizado para ubicar puntos coordinados máximos de la región en la matriz(imagen).
- Multiplicación.
- Dilatación.
- Erosión.

Las tres últimas técnicas fueron muy utilizadas en este algoritmo. Serán nombradas pero no explicadas, debido a que ya hemos hablado de ellas anteriormente.

Cuando se aplica la técnica de umbralización de Ridler Calvard, es en realidad una modificación de ella, ya que en primer lugar se calcula el valor medio de las intensidades de los píxeles y luego se divide el vector general en dos vectores (uno cuyos valores son menores que el umbral y el otro al contrario). Comenzamos la interacción con el vector superior, por lo tanto el umbral encontrado resultará mayor al que encontraría el método aplicado directamente a la imagen total.

Veremos en las siguientes imágenes el resultado de este método y la ubicación de los puntos generadores para la etapa siguiente, refiriéndonos al crecimiento de región.

El criterio con el cual se han ubicado estos puntos también fue propuesto luego de varias pruebas, pero la idea principal es de tomarlos automáticamente en la esquina superior derecha si la imagen tiene Laterality = “R”, o en la esquina superior izquierda si la misma tiene Laterality = “L”, siempre y cuando estén dentro de la región blanca de la imagen binarizada.

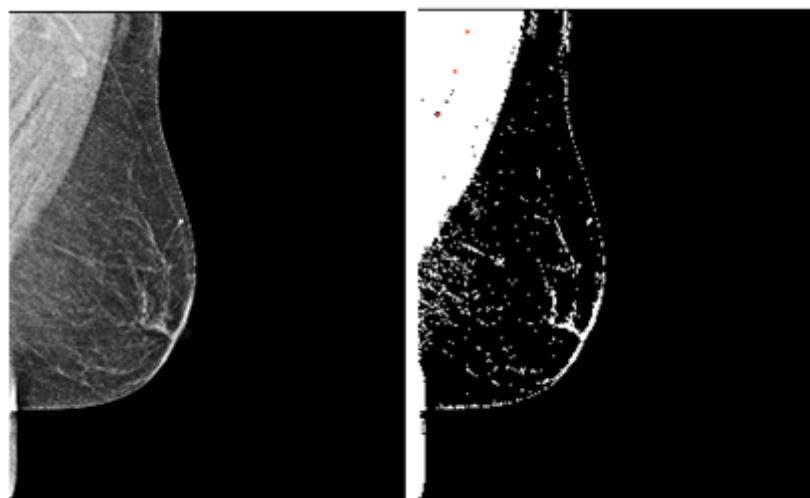


Figura 3.4.3: Umbralización y ubicación de puntos generadores (“Semillas”).

Una vez posicionados los puntos se procede a aplicar el crecimiento de región a la imagen resultante de las etapas anteriores, especificando un determinado umbral (umbral obtenido luego de aplicar Ridler Calvard).

El operador gradiente se utilizó para aumentar el contraste en los bordes dentro de la imagen. Luego se realiza una suma ponderada entre el y la imagen sobre la cual se le aplicará el crecimiento de región.

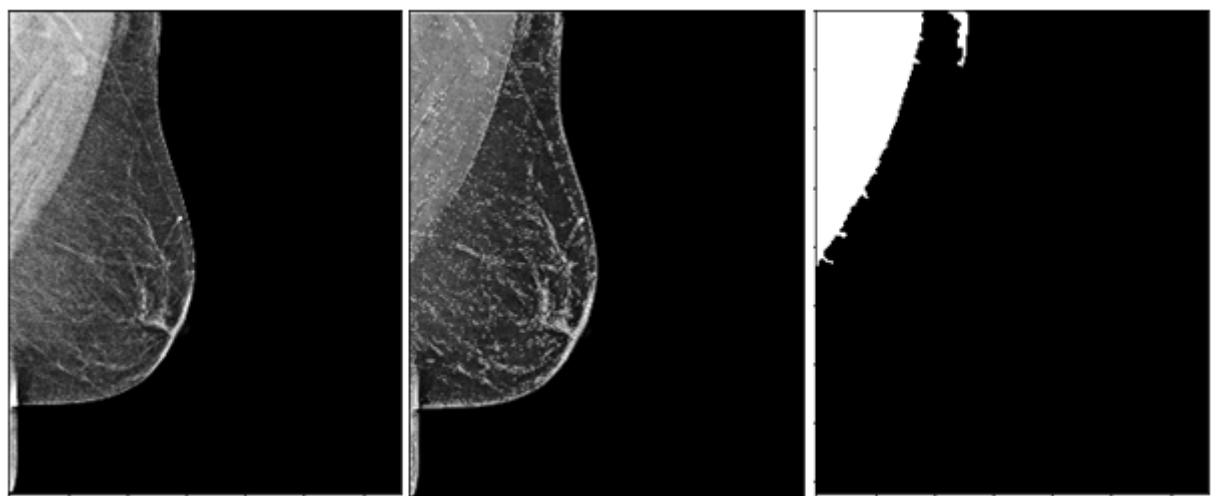


Figura 3.4.4: Aplicación de operador gradiente y crecimiento de región.

Por último se realiza un ajuste de curva, aproximando las coordenadas de los puntos tomados a un polinomio de segundo orden, ya que el músculo posee una curvatura similar.

Coordenadas en x:

[300. 500. 700. 900. 1000. 1100. 1200. 1300. 1400. 1600. 1700. 1900.
2000. 2075.]

Coordenadas en y:

[944. 936. 841. 811. 729. 612. 568. 476. 394. 358. 309. 167. 51. 0.]

Figura 3.4.5: Coordenadas de los puntos para el ajuste de curva.

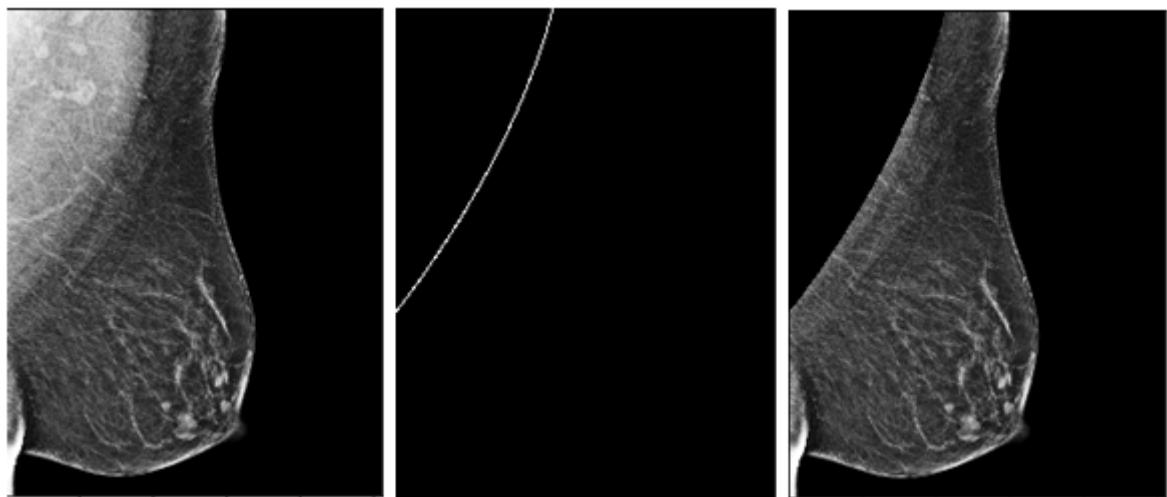


Figura 3.4.6: Curva obtenida e imagen resultante.

En caso de que el resultado no sea el correcto, con las mismas coordenadas se realiza otro ajuste pero ahora de una recta y se vuelve a corroborar automáticamente. Explicaremos luego el criterio con el cual se evalúa la curva.

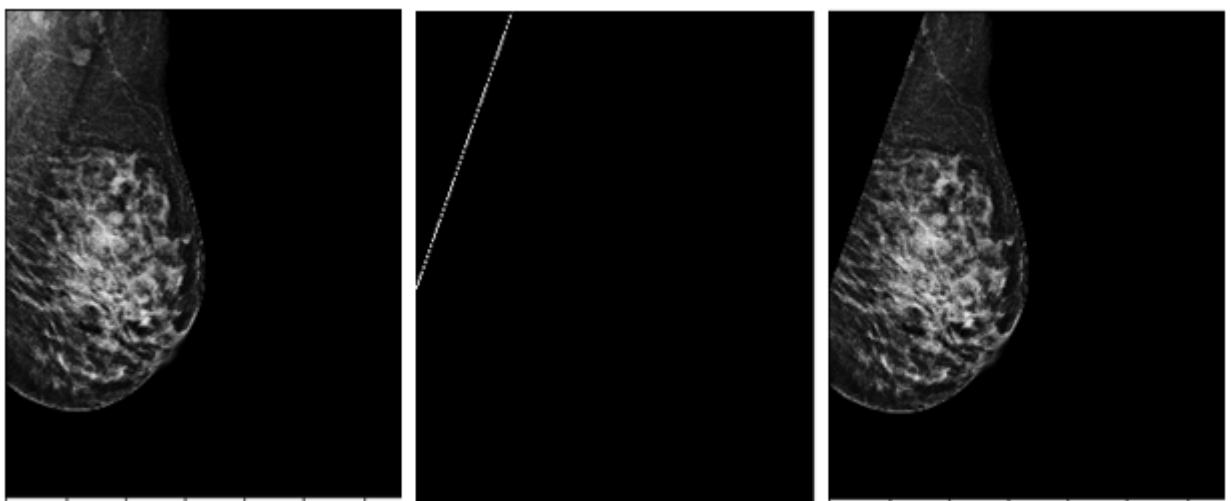


Figura 3.4.7: Recta obtenida e imagen resultante.

El algoritmo de evaluación de la curva determina:

- Convergencia de la misma: Siendo la variable independiente X (coordenada correspondiente a la fila), se verifica que los valores de Y (variable dependiente) obtenidos vayan disminuyendo hasta llegar a cero (columna cero) si la imagen tiene Laterality = “L”, o vayan aumentando hasta llegar a la coordenada de máxima columna si Laterality = “R”. Necesitamos que converja debido a que el paso posterior es realizar un crecimiento de región, y ya que éste necesita un límite, la curva no se puede cortar.
- Coordenadas del punto inicial y final: tomando como coordenada inicial el punto donde X=0 e Y es igual a algún valor, si la imagen tiene Laterality = “L”, el mismo tiene que ser mayor a $\frac{columnamax}{2}$ siendo columnamax la coordenada de columna donde termina la región de la imagen que posee valores superiores a cero con coordenada fila igual a cero, y si tomamos como coordenada final el par ordenado donde Y=0 y X es igual a algún valor, el mismo tiene que ser mayor a $\frac{filamax}{2}$, refiriéndonos a filamax como última fila donde la imagen tiene valores superiores a cero y no a la fila máxima correspondiente al tamaño de la imagen. Si la imagen tiene Laterality = “R”, el valor de Y en la coordenada inicial donde X=0, tiene que ser menor a $\frac{(columnamax - columnamin)}{2} + columnamin$ debido a que la mama está posicionada en el borde donde la columna es máxima y no cero. El análisis es similar para la coordenada final donde Y=columnamax, X tiene que ser mayor a $\frac{filamax}{2}$. Esto se ha tomado en consideración luego de visualizar las imágenes y realizar pruebas.
- Por lo mencionado anteriormente se espera que la curva converja luego de $\frac{filamax}{2}$, pero ésta debe realizarse antes de filamax, ya que el músculo termina en algún momento antes de que termine la región donde los píxeles son distintos de cero. Por lo que se propone como límite máximo de convergencia $\frac{3}{4}filamax$, si esto no ocurre el algoritmo pone en alto la bandera que indica que se ajuste la recta.

Esta verificación de convergencia y posición es muy importante para que el algoritmo sea mas robusto y pueda adaptarse a los distintos casos con un óptimo resultado.

Generalmente se ajusta una recta en los casos en donde el músculo es muy poco intenso o su límite no está bien marcado, ya que se confunde con el tejido glandular contiguo (mayormente en imágenes de densidad “d”). También sucede esto cuando el crecimiento de región no funciona bien (crece más o menos de lo que debería), como ocurrió en la figura 3.4.4.

Luego de obtener la recta se realiza el mismo procedimiento para verificar si es correcta. Si no pasa la prueba de convergencia se vuelve a ajustar una recta pero sin tener en cuenta los puntos de coordenadas X e Y como se muestran en la figura 3.4.5, Se definen nuevas coordenadas (punto inicial y final) como: $(0, \frac{3}{4}columnamax)$, $(\frac{3}{4}filamax, 0)$ si la imagen tiene Laterality = “L”, o $(0, \frac{1}{2}(columnamax - columnamin) + columnamin)$, $(\frac{3}{4}filamax, columnamax)$ si la imagen tiene Laterality = “R”.

A continuación se darán mas ejemplos donde se han ajustado curvas o rectas.

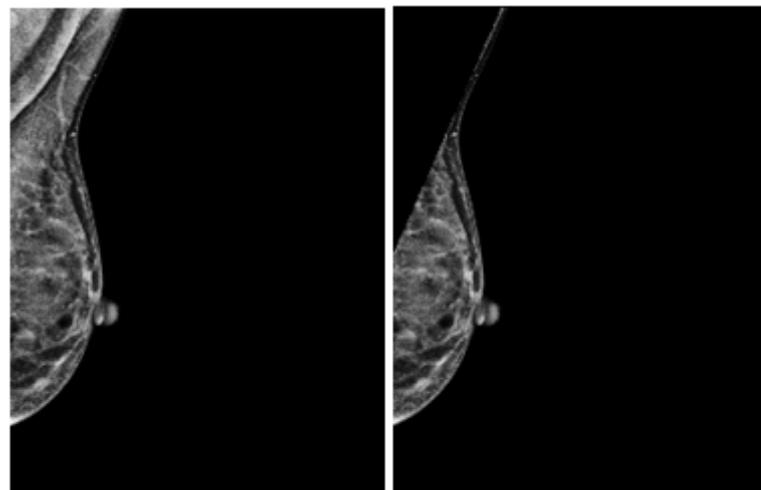


Figura 3.4.8: Músculo se confunde con glándula.

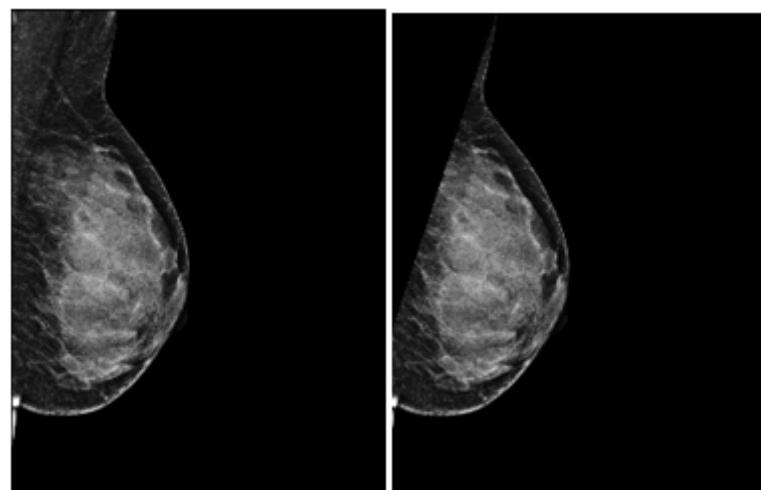


Figura 3.4.9: Músculo poco intenso.

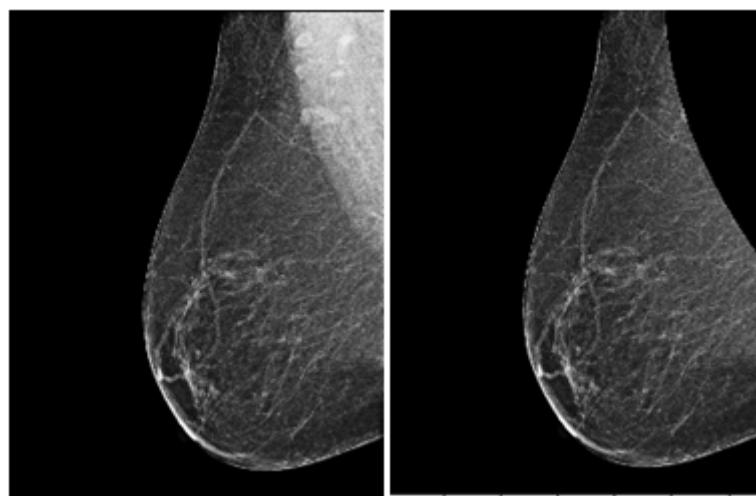


Figura 3.4.10: Músculo bien definido.

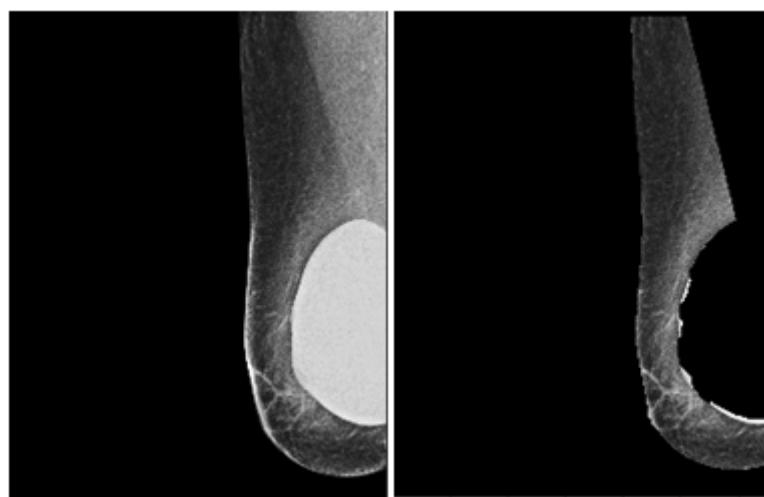


Figura 3.4.11: Extracción de músculo a imagen con implante mamario.

3.4.4. Sustracción de piel

Esta etapa es muy simple ya que se aplica a todas las imágenes por igual. Consiste en eliminar el borde de la región de interés obtenida, evitando que quede muy intenso como ocurre en algunos casos.

Las técnicas implementadas son:

- Binarización.
- Contorno.
- Dilatación.
- Negativo.

- Multiplicación.

Los pasos que sigue este algoritmo se demostrarán en la siguiente sucesión de imágenes:

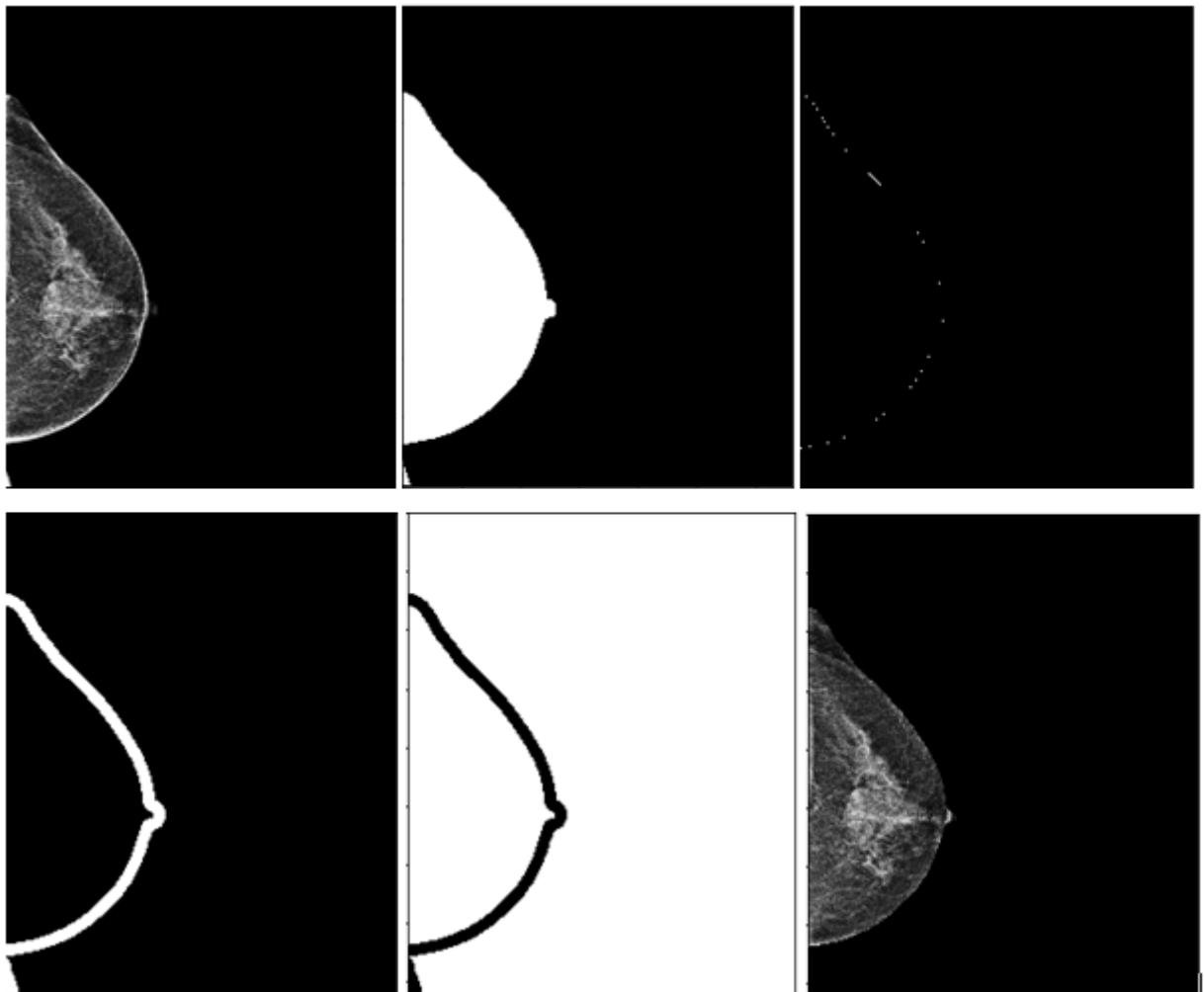


Figura 3.4.12: Sustracción de piel.

Detalles debido al ruido de la imagen también pueden ser eliminados, como ocurre en la figura anterior (esquina inferior izquierda).

Una vez construido el algoritmo de segmentación con todas sus adaptaciones correspondientes, se procedió a hacer el recorrido para todas las imágenes de nuestra base.

Los resultados obtenidos fueron óptimos en 99 % mamografías segmentadas. Las restantes necesitaron de una mínima intervención para evitar arrastrar errores en las siguientes etapas. Esto se logró modificando los siguientes parámetros:

- Coordenadas de semillas utilizadas en el crecimiento de región, tanto para músculo y/o implante.
- Coordenadas de punto inicial y final para el ajuste de curva.

3.5. Extracción de características

En esta sección se explicará como fueron desarrollados los algoritmos de extracción de características (brillo y textura) de la región de interés obtenida anteriormente. Las mismas formarán parte de la base de datos utilizada en la etapa de clasificación. Ellas son:

- Cuantificación de píxeles de la ROI.
- Análisis de histograma global.
- Análisis de histograma local.
- Entropía de permutaciones.
- Entropía de permutaciones (aplicada a imagen umbralizada).
- Análisis fractal.
- Descriptores de Haralick.
- Características extraídas de información DICOM.

3.5.1. Cuantificación de píxeles de la ROI

Luego de la visualización de imágenes de distintas clases, se ha notado que las mujeres que poseen mamas de mayor tamaño, en general, presentan una proporción de tejido graso superior.

Por esta razón tomamos como una característica la cantidad de píxeles que componen la región de interés como medida del área de la mama. Debe tenerse en consideración el tamaño del píxel (resolución de la imagen). En nuestro caso, todas las imágenes fueron adquiridas con la misma resolución, por lo que no se requirió de ninguna operación extra para la obtención de este atributo.

Esto se realiza binarizando la imagen y contando aquellos píxeles con valor igual a uno.

3.5.2. Análisis de histograma global

Una vez obtenida la ROI, se procede a trabajar con la imagen como una matriz de datos (niveles de grises de cada píxel), de la cual solo nos interesa la región de primer plano (región donde los datos son mayores a cero) correspondiente al tejido mamario.

Como en el análisis de histograma global no interesa la posición de los píxeles, se generará un vector cuyo contenido corresponde a los niveles de grises mayores a cero.

Luego se realiza la gráfica representando en el eje de las abscisas los niveles de grises, que en nuestras imágenes van desde 0 hasta 65536 (2^6); aunque el valor máximo ronda los 4000 para la mayoría; y en el eje de las ordenadas la cantidad de píxeles correspondientes a cada nivel.

Las características estadísticas extraídas de este histograma son:

- Media.
- Mediana.
- Desviación estándar.
- Asimetría.
- Curtosis.
- Moda.

A modo de ejemplo se mostrarán los valores de las mismas para cuatro imágenes. Mamografía etiquetada con densidad a, b, c y d respectivamente.

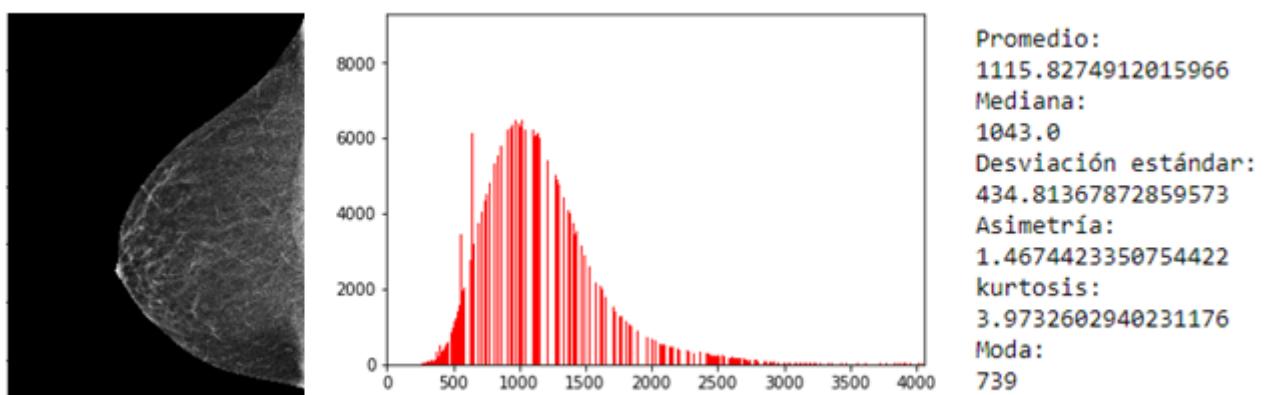


Figura 3.5.1: Histograma de mamografía con densidad a.

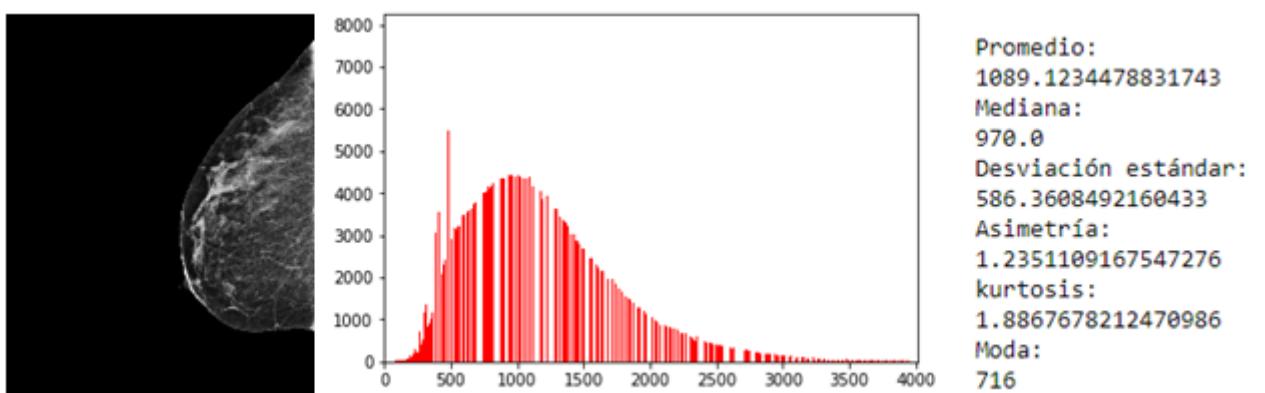


Figura 3.5.2: Histograma de mamografía con densidad b.

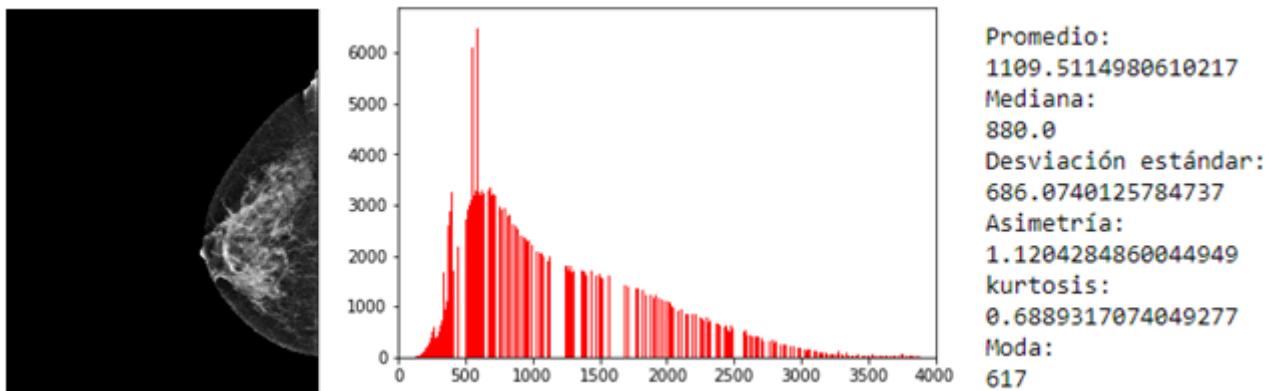


Figura 3.5.3: Histograma de mamografía con densidad c.

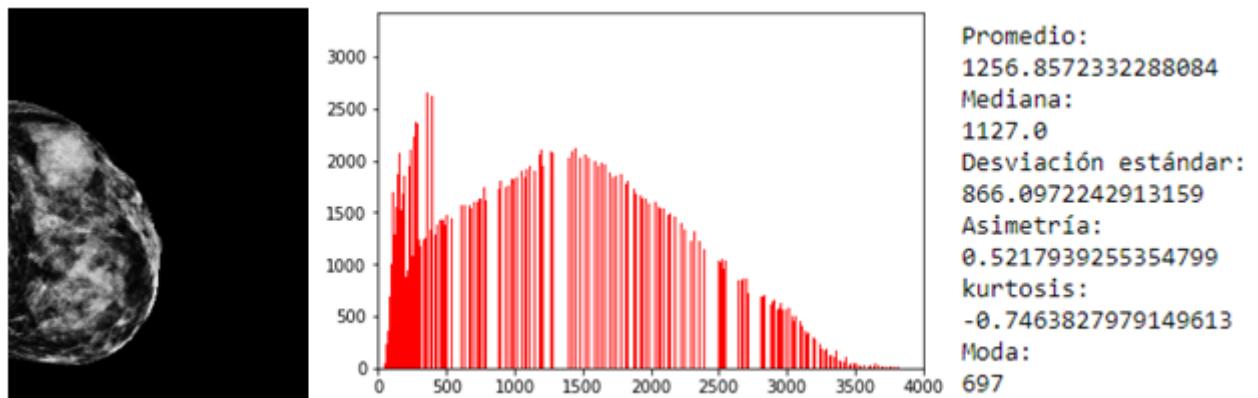


Figura 3.5.4: Histograma de mamografía con densidad d.

Se pueden plantear las siguientes hipótesis:

- La desviación estándar va aumentando a medida que aumenta la densidad (datos con mayor dispersión alrededor de la media).
- La asimetría va disminuyendo a medida que aumenta la densidad (aumenta la simetría de la curva).
- La curtosis va disminuyendo a medida que aumenta la densidad (va disminuyendo el grado de apuntamiento de la distribución).

3.5.3. Análisis de histograma local

Las características basadas en histogramas de nivel de gris, como desviación estándar y asimetría, resultaron ser prometedoras.

Un problema grave con éstas es que dependen de los parámetros de adquisición de las imágenes y que son muy sensibles a los cambios en el grosor del tejido en la periferia de la mama.

Para tratar la variación del grosor del tejido en la periferia, se investigó el uso de características basadas en histogramas realizados en función de la distancia a la línea de la piel. Esta idea se basa en el supuesto de que existirá una fuerte correlación entre el grosor del tejido y la distancia, ya que el seno se comprime entre dos placas rígidas durante el procedimiento de adquisición, y la periferia del seno se compone principalmente de grasa que es muy flexible.

Basado en el estudio realizado en [35] se procedió a reproducir la idea. Esta etapa consiste en segmentar la mama en 5 partes diferentes (a un 20 %, 40 %, 70 % y 90 % del borde de la mama, respectivamente), luego se calculan los valores de desviación estándar y asimetría de los histogramas de las 3 partes centrales (la del 40 %, el 70 % y el 90 %) y se los utiliza como vectores de características.

La división de regiones según la distancia, se realizó aplicando:

- Binarización.
- Erosión.
- Multiplicación.

El elemento estructurante utilizado para erosionar la imagen es calculado según el porcentaje que se requiere desde el borde. A continuación se demuestra el procedimiento.

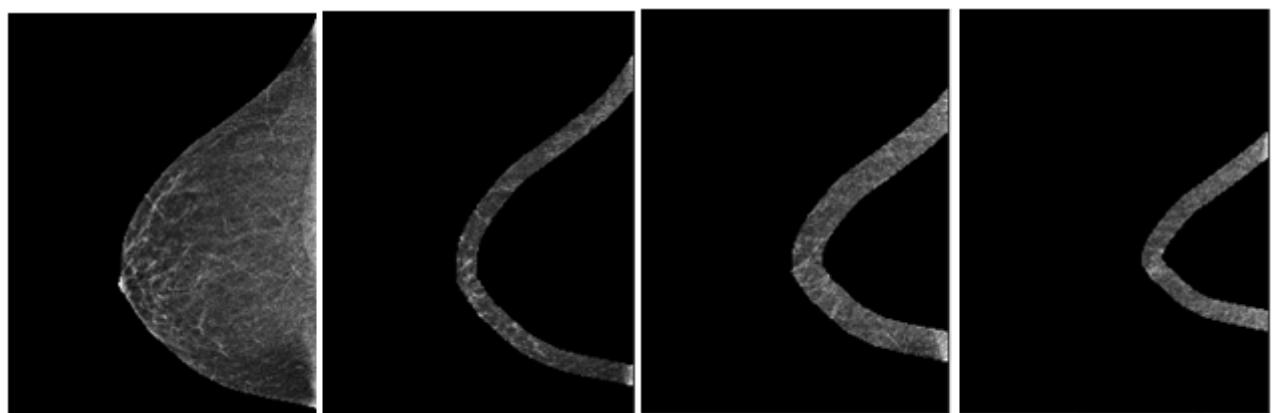


Figura 3.5.5: División de la mama en regiones según la distancia a la piel.

De este análisis se obtienen dos vectores, uno constituido por los valores de desviación estándar de las tres porciones que se tuvieron en cuenta, y el otro compuesto por los valores de asimetría.

```
Desviación 20-40/40-70/70-90 % desde piel:  
[364.98112441 320.20483162 346.46089778]  
Asimetría 20-40/40-70/70-90 % desde piel:  
[1.5440883 1.04769628 1.17396022]
```

Figura 3.5.6: Vectores obtenidos para mamografía de densidad a.

```
Desviación 20-40/40-70/70-90 % desde piel:  

[416.09585325 478.2033173 448.64612554]  

Asimetría 20-40/40-70/70-90 % desde piel:  

[1.63816364 0.97570646 0.88043264]
```

Figura 3.5.7: Vectores obtenidos para mamografía de densidad b.

```
Desviación 20-40/40-70/70-90 % desde piel:  

[695.75299574 768.9876474 638.01612525]  

Asimetría 20-40/40-70/70-90 % desde piel:  

[1.78857684 0.77929491 0.36972212]
```

Figura 3.5.8: Vectores obtenidos para mamografía de densidad c.

```
Desviación 20-40/40-70/70-90 % desde piel:  

[932.66328139 849.609252 753.52727402]  

Asimetría 20-40/40-70/70-90 % desde piel:  

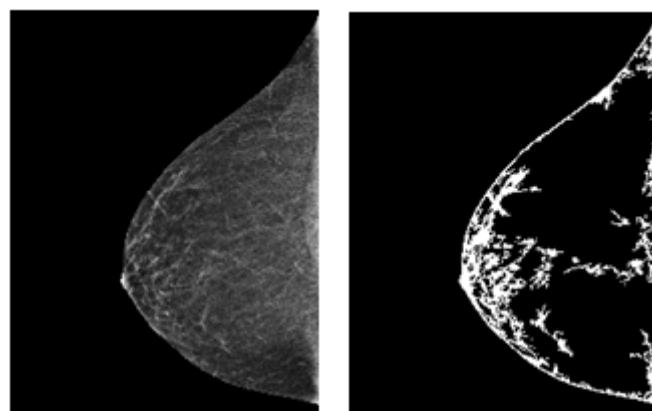
[0.92294062 0.26152098 0.24910378]
```

Figura 3.5.9: Vectores obtenidos para mamografía de densidad d.

De estos ejemplos se puede observar cómo los valores de desviación van aumentado con la densidad, y por el contrario los de asimetría disminuyen.

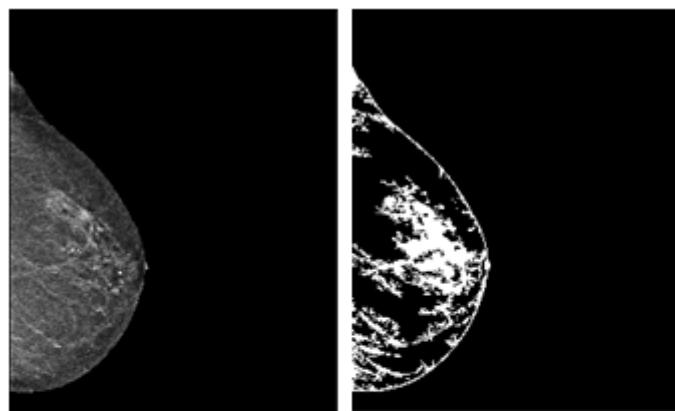
Cuando se aplica el operador gradiente a una imagen, éste devuelve otra que puede ser analizada de la misma manera. Si a la imagen resultante se umbraliza (utilizando un umbral fijo), se notará que las mamografías correspondientes a la clase A tienen pocos píxeles blancos en comparación al total. Por lo tanto esa relación entre cantidad de píxeles blancos y totales se la extrae como una característica más.

Por otro lado luego de visualizar las imágenes umbralizadas, se observó que si se las divide en porciones como en el análisis anterior y se calculan las relaciones entre píxeles blancos y totales de esa región particular, se obtendrán valores que pueden resultar significativos para la clasificación. A su vez, se añade una característica que representa el promedio de niveles de grises de la imagen gradiente.



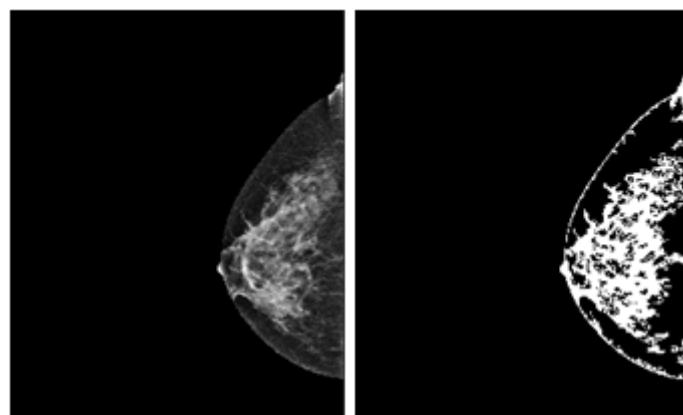
Promedio de imagen con Grandiente:
17.219883
Cant. pixeles blancos/ Cant. pixeles totales:
13.12770188257024
Relación 20/40/70/90 desde piel:
[29.8531749 12.57523411 4.65600916 4.10255533]

Figura 3.5.10: Relaciones obtenidas para mamografía de clase a.



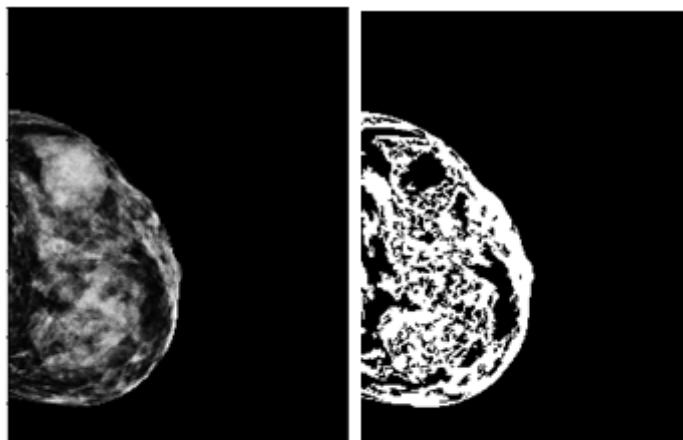
Promedio de imagen con Grandiente:
20.561337
Cant. pixeles blancos/ Cant. pixeles totales:
21.73308619091752
Relación 20/40/70/90 desde piel:
[19.67371889 20.99162436 28.17754933 25.90273737]

Figura 3.5.11: Relaciones obtenidas para mamografía de clase b.



Promedio de imagen con Grandiente:
22.346212
Cant. pixeles blancos/ Cant. pixeles totales:
35.07335266970066
Relación 20/40/70/90 desde piel:
[22.19284261 29.5781056 38.79712711 44.1889144]

Figura 3.5.12: Relaciones obtenidas para mamografía de clase c.



Promedio de imagen con Grandiente:
27.022635
Cant. pixeles blancos/ Cant. pixeles totales:
48.159851775633236
Relación 20/40/70/90 desde piel:
[63.4811048 32.93765818 40.88227144 52.07667949]

Figura 3.5.13: Relaciones obtenidas para mamografía de clase d.

En esta etapa se pueden plantear las siguientes hipótesis:

- El promedio de la imagen gradiente aumenta a medida que aumenta la densidad (mayor cambios en la textura).
- Las relaciones entre cantidad de píxeles blancos y totales aumenta con la densidad.

3.5.4. Entropía de permutaciones

La idea de utilizar la entropía de permutación como una medida de densidad mamaria fue propuesta en [31]. En este caso no se procedió a reproducir el trabajo, debido que éste plantea diversos modelos, pero sí se ha realizado un cálculo para luego utilizarlo como característica.

Una vez obtenida la ROI, se la dividirá en distintas porciones, refiriéndonos a ellas como celdas de 100x100 píxeles. Luego a cada una se le calcula la entropía de permutaciones con los siguientes parámetros:

- t=1 (retardo).
- d=6 (dimensión de inmersión).

El procedimiento es el siguiente:

- Se obtiene la celda de 100x100 píxeles.
- Se posiciona en la esquina izquierda superior una matriz 3x2.
- Se toman los valores de nivel de gris de los píxeles encerrados por esa matriz.
- Se calcula y se guarda el vector de permutaciones (en este caso habría $6!$ combinaciones posibles).
- Se corre un lugar hacia la derecha la matriz (ya que el retardo es uno) y se obtiene el nuevo vector de permutaciones. Se repite hasta recorrer toda la celda.
- Se calcula la proporción de aparición de cada combinación.
- Se calcula la entropía de permutaciones de esa celda.
- Se toma otra celda y así sucesivamente hasta recorrer toda la ROI.
- El resultado final corresponde al promedio de entropía de todas las celdas.

Luego de aplicar este algoritmo a varias imágenes se ha notado que no hay mucha variación en los valores obtenidos según sea la clase de mamografía. Sin embargo, de este análisis surgió la idea de calcular la entropía de permutaciones a la imagen que resulta de aplicar el gradiente y umbralización como se muestra en la Figura 3.5.11. Por lo que ésta constituye una característica de la base construida luego.

Otro parámetro que se analizó fue la relación entre la cantidad de celdas en las cuales su entropía resultó igual a cero y la cantidad de celdas totales analizadas. Esto se puede ver a continuación:

```

Entropía de permutación
0.7736587111091129

Entropía de permutación de imagen binaria
0.0425979221657674
Cant. cuadrados con entropía cero/ Cant. cuadrados totales:
0.6993006993006993

```

Figura 3.5.14: Entropías calculadas para mamografías de clase a.

```

Entropía de permutación
0.7723367937407187

Entropía de permutación de imagen binaria
0.08197710708816337
Cant. cuadrados con entropía cero/ Cant. cuadrados totales:
0.47368421052631576

```

Figura 3.5.15: Entropías calculadas para mamografías de clase b.

```

Entropía de permutación
0.7696798285923908

Entropía de permutación de imagen binaria
0.1137194450782655
Cant. cuadrados con entropía cero/ Cant. cuadrados totales:
0.30414746543778803

```

Figura 3.5.16: Entropías calculadas para mamografías de clase c.

```

Entropía de permutación
0.7657661562900615

Entropía de permutación de imagen binaria
0.15685951954763758
Cant. cuadrados con entropía cero/ Cant. cuadrados totales:
0.08673469387755102

```

Figura 3.5.17: Entropías calculadas para mamografías de clase d.

El procedimiento para realizar el cálculo de entropía de permutaciones de la imagen binaria fue similar al explicado anteriormente, con la diferencia que se utilizó una matriz de 3x3 (cantidad de combinaciones posibles: 2^9) ya que los valores son ceros o unos. Cabe aclarar que anteriormente se utilizó una matriz 3x2 para disminuir el costo computacional ($9!$ es mucho mayor a $6!$).

Hipótesis planteadas:

- La entropía de permutaciones calculada en la imagen binaria aumenta levemente a medida que aumenta la densidad.

- La relación entre celdas con entropía cero y totales disminuye con el aumento de densidad.

3.5.5. Análisis fractal

El procedimiento adoptado para el cálculo de la dimensión fractal fue el siguiente:

1. Obtención, por medio del algoritmo multinivel de Otsu, un set de cuatro umbrales $t_i \in T$.
2. Descomposición de la imagen de entrada en escala de grises $I(x, y)$ en una serie de imágenes binarias $I_b(x, y)$. Para ello, al set de umbrales T de Otsu se le agrega n_m , que corresponde al máximo nivel de gris en $I(x, y)$, luego se toma de a dos umbrales t_i y se aplica:

$$I_b(x, y) = \begin{cases} 1 & \text{si } t_i < I(x, y) \leq t_s \\ 0 & \text{en otros casos} \end{cases} \quad (3.1)$$

Donde t_i y t_s denotan el umbral inferior y superior respectivamente. Éstos serán un par de umbrales consecutivos $\{t_i, t_{i+1}\}$ y cada $t_i \in T$ con $n_m: \{t_i, u_l\}$. De esta manera, si la cantidad de umbrales era cuatro, obtendremos 2^4 imágenes binarias (ver Figura 3.5.18).

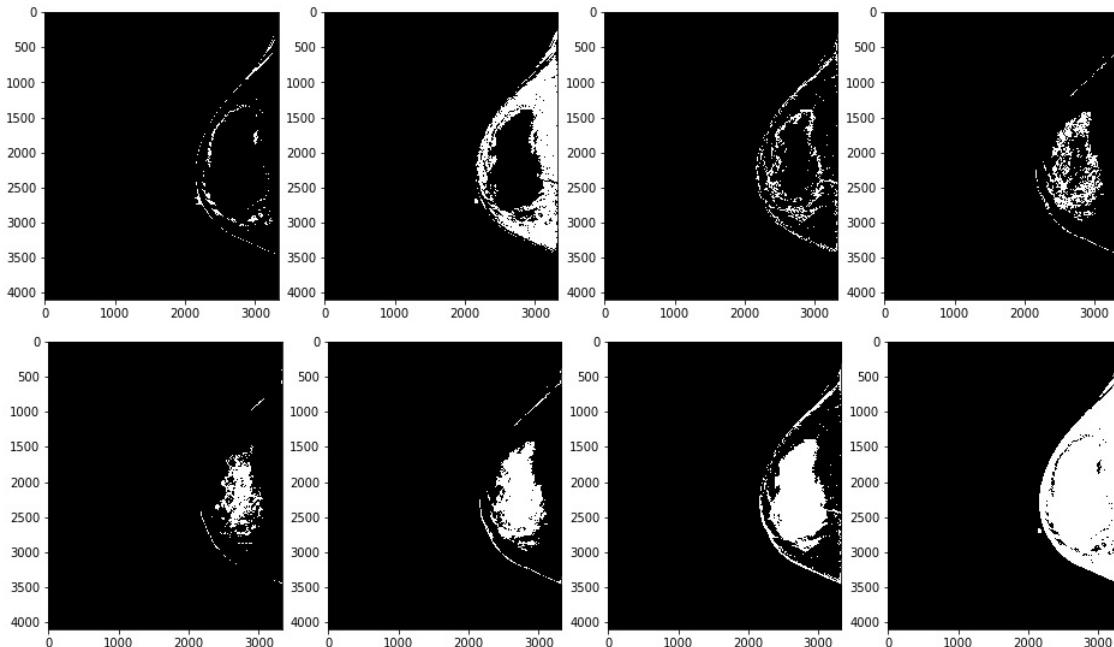


Figura 3.5.18: Imágenes segmentadas por medio del algoritmo de Otsu multiumbral.

3. Aproximación de la dimensión fractal para cada una de las imágenes binarias aplicando el algoritmo de conteo de recuadros. Como se mencionó anteriormente, el modelo será una buena aproximación si los datos se ajustan a una recta por medio de una regresión lineal. Como podemos ver en la Figura 3.5.19 esto se cumple muy bien para nuestras imágenes. La pendiente de dicha recta será el valor de la dimensión fractal.

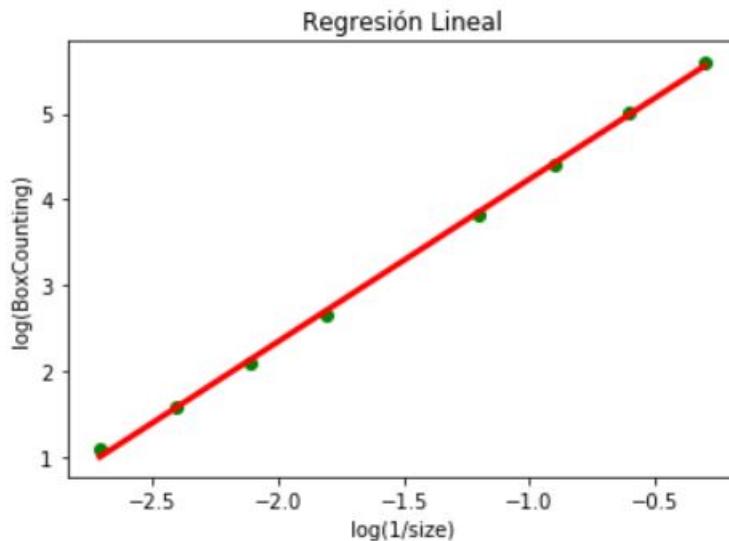


Figura 3.5.19: Ajuste de recta en algoritmo de conteo de recuadros (Box Counting).

4. El vector de características de la dimensión fractal estará acompañado de otros valores (ver en Figura 3.5.20), éstos son:

- Área (cantidad de píxeles que forman la imagen binaria).
- Promedio de nivel de gris. Esto es el promedio de nivel de gris en $I(x, y)$ en el área delimitada por la segmentación que corresponda.
- Desvío estándar de los valores de grises.
- Curtosis.
- Asimetría.
- Entropía.

| Image | FD | Area | Mean grey level | Standard Deviation | Kurtosis | Skewness | Entropía |
|---------------|---------|-----------|-----------------|--------------------|-----------|-------------|----------|
| binaryImage_1 | 1.41731 | 0.0498363 | 158.416 | 58.8302 | -0.272962 | -0.851232 | 3.96946 |
| binaryImage_2 | 1.84299 | 0.521263 | 456.59 | 135.241 | -0.520876 | 0.537627 | 4.92633 |
| binaryImage_3 | 1.60771 | 0.133691 | 1123.13 | 224.053 | -1.28544 | 0.199953 | 4.0968 |
| binaryImage_4 | 1.65526 | 0.16334 | 1934.76 | 224.393 | -1.18012 | -0.0858978 | 4.31008 |
| binaryImage_5 | 1.62971 | 0.128002 | 2664.59 | 247.531 | -0.315378 | 0.63404 | 4.37622 |
| binaryImage_6 | 1.72766 | 0.291792 | 2255.5 | 431.358 | -0.770175 | 0.262938 | 4.68502 |
| binaryImage_7 | 1.78589 | 0.425834 | 1899.39 | 647.651 | -1.00108 | -0.00469737 | 4.73915 |
| binaryImage_8 | 1.94681 | 0.947547 | 1105.15 | 844.69 | -0.702123 | 0.837583 | 4.55267 |

Figura 3.5.20: Vector con el cálculo de la dimensión fractal y de sus valores asociados.

También se decidió realizar otra variante de este método [24], en la que se obtuvo la dimensión fractal de los bordes de las imágenes binarias. En lo único que difiere respecto al procedimiento

anterior es que entre los items 2 y 3 se buscaron los bordes de las imágenes, tal como se ilustra en las Figuras 3.5.21 y 3.5.22.

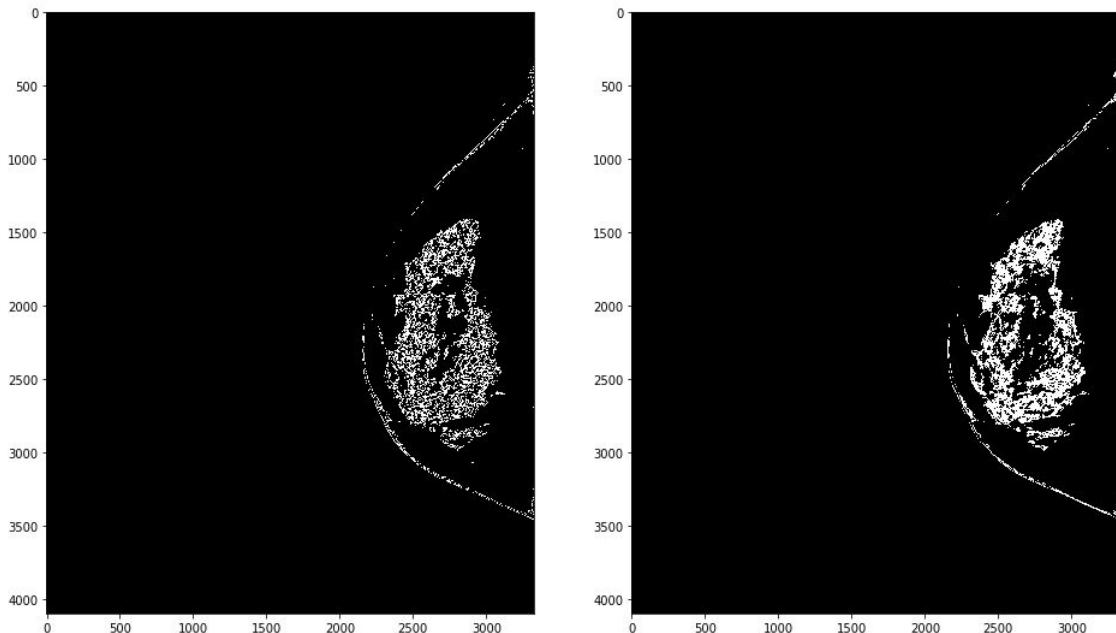


Figura 3.5.21: Bordes de la imagen segmentada.

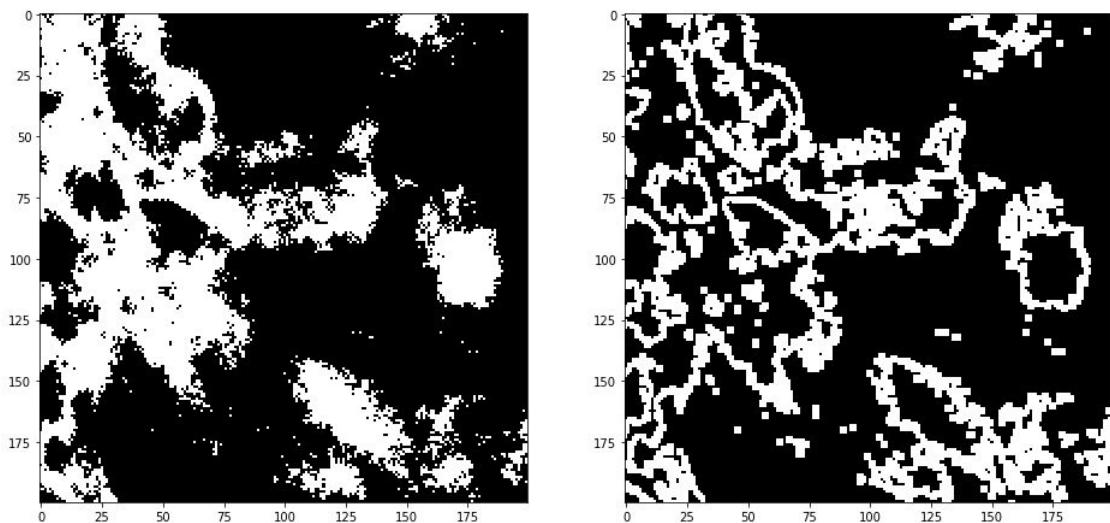


Figura 3.5.22: Aumento de los bordes de la imagen segmentada.

Finalmente se obtuvo un vector de las mismas características que el anterior.

Hipótesis planteadas:

- Cuando hay un alto grado de densidad mamográfica, la imagen aparecerá más suave, lo que debería reflejarse en una dimensión fractal más baja.
- Una imagen de un seno con un pequeño grado de densidad tendrá una textura más grosera,

debido al buen contraste entre el tejido conectivo y el tejido glandular predominantemente graso, por lo cual se espera que la dimensión fractal sea mayor.

3.5.6. Descriptores de textura de Haralick

En este caso, simplemente se utilizó un comando de la librería mahotas de Python: `mahotas.features.haralick()`. De este modo se calculan trece de las catorce características de Haralick [25] en las cuatro direcciones de las imágenes bidimensionales, obteniendo una matriz de dimensión 4×13 .

```
Descriptores de Haralick

[[ 1.15891720e-05  7.91443643e+03  9.94640226e-01  7.38318068e+05
   2.78092180e-02  2.12945404e+03  2.94535784e+06  1.20560741e+01
   1.80121574e+01  1.99532422e-06  7.38577170e+00 -2.96980890e-01
   9.99064826e-01]
 [ 9.86674241e-06  1.07282157e+04  9.92720843e-01  7.36913343e+05
   2.40512368e-02  2.12889729e+03  2.93692516e+06  1.20561448e+01
   1.82300091e+01  1.69226751e-06  7.60660437e+00 -2.76350617e-01
   9.98552640e-01]
 [ 1.15228407e-05  7.75738242e+03  9.94744782e-01  7.38064785e+05
   2.84036867e-02  2.12891826e+03  2.94458176e+06  1.20564622e+01
   1.80055161e+01  2.00270279e-06  7.37691880e+00 -2.97699426e-01
   9.99079248e-01]
 [ 1.00050855e-05  1.05188910e+04  9.92862198e-01  7.36843856e+05
   2.51816502e-02  2.12840782e+03  2.93685653e+06  1.20561473e+01
   1.81932197e+01  1.73855832e-06  7.58340216e+00 -2.79900676e-01
   9.98657748e-01]]
```

Figura 3.5.23: Ejemplo de la matriz de los descriptores de Haralick.

3.5.7. Características obtenidas de la información DICOM

De la información contenida en los archivos DICOM se extrajo las siguientes características[14]:

- *Edad del paciente*
- *Vista del estudio:* Puede tomar tres valores diferentes:
 - CC :Vista craneo caudal de la imagen.
 - MLO: Vista oblicua mediolateral.
 - IDMLO:Vista oblicua mediolateral en pacientes con prótesis.
- *Lateralidad:* Puede tomar dos valores:
 - R: derecha.
 - L: izquierda.
- *KVP:* Kilo voltaje pico de salida del generador.
- *Corriente del tubo de rayos X,* medida en mili Amperes.

- *Exposición:* Corriente del tubo de rayos X medida en mili Amperes por segundo.
- *Tiempo de exposición:* Duración de la exposición a los rayos X medida en mili segundos.
- *Material del filtro:*
 - Plata
 - Aluminio
 - Molibdeno
 - Rodio
- *Espesor del filtro*
- *Punto focal:* Tamaño nominal del punto focal en mm utilizado para adquirir la imagen.
- *Material del blanco del ánodo:* Puede estar construido de:
 - Tungsteno
 - Molibdeno
 - Rodio
- *Espesor de la mama:* El grosor promedio en mm de la parte del cuerpo examinada cuando se comprime, si se ha aplicado compresión durante la exposición.
- *Fuerza de compresión:* La fuerza de compresión aplicada a la parte del cuerpo durante la exposición, medida en Newtons.
- *Distancia de la fuente al detector:* Distancia en mm desde la fuente hasta el centro del detector.
- *Distancia de la fuente al paciente*
- *Estimación de la magnificación*
- *Angulación primaria del posicionador:* Distancia en mm desde la fuente hasta la mesa, el lado de apoyo más cercano al sujeto de imagen, medido a lo largo del rayo central del haz de rayos X.
- *Modo de control de la exposición:* Existen tres posibles configuraciones:
 - Automático.
 - Autofiltrado.
 - Manual.
- *Descripción del modo de control de la exposición*

- *Exposición relativa a los rayos X.*
- *Dosis de entrada:* El valor medio de la dosis de entrada medido en mili Gray en la superficie del paciente durante la adquisición de esta imagen.
- *Dosis que recibe el órgano expuesto a la radiación:* Valor medio de la dosis en el órgano medido en deci Gray durante la adquisición de esta imagen.
- *Capa de valor medio:* Grosor del aluminio en mm requerido para reducir la salida de rayos X en un factor de dos.
- *Presencia de implante mamario:* Puede tomar dos valores:
 - Sí.
 - No.

3.6. Preprocesamiento de datos

El preprocesamiento se refiere a las transformaciones aplicadas a los datos antes de incluirlos en los algoritmos de aprendizaje.

Los distintos algoritmos de clasificación tienen afinidad con ciertos tipos de datos y de ello depende su desempeño. Es de allí la gran importancia de que los datos reciban un tratamiento adecuado. A continuación se comentará cómo se trabajó.

3.6.1. Análisis de características

En primer lugar, se realizó un análisis exploratorio de las características. Para ello se hizo uso de los diagramas de caja y bigotes que nos proporcionan una visión general de la simetría de la distribución de los datos.

Además, este tipo de gráficos son útiles para ver la presencia de valores atípicos (outliers). Pertenece a las herramientas de la estadística descriptiva. Permite ver como es la dispersión de los puntos con la mediana, los percentiles 25 y 75 y los valores máximos y mínimos.

1. Análisis de las características del histograma

A continuación se procederá a analizar conjuntamente las características, tanto del histograma global como del local.

En las Figuras 3.6.1 y 3.6.2 se muestran los gráficos de caja correspondientes al desvío estándar y la kurtosis respectivamente. Se evidencia que las hipótesis planteadas anteriormente se cumplen. Esto es, el desvío estándar se incrementa con el aumento de la densidad mamaria mientras que la curtosis disminuye. No es así para el caso de la asimetría (ver Figura 3.6.3), donde su variabilidad con la densidad no es tan marcada.

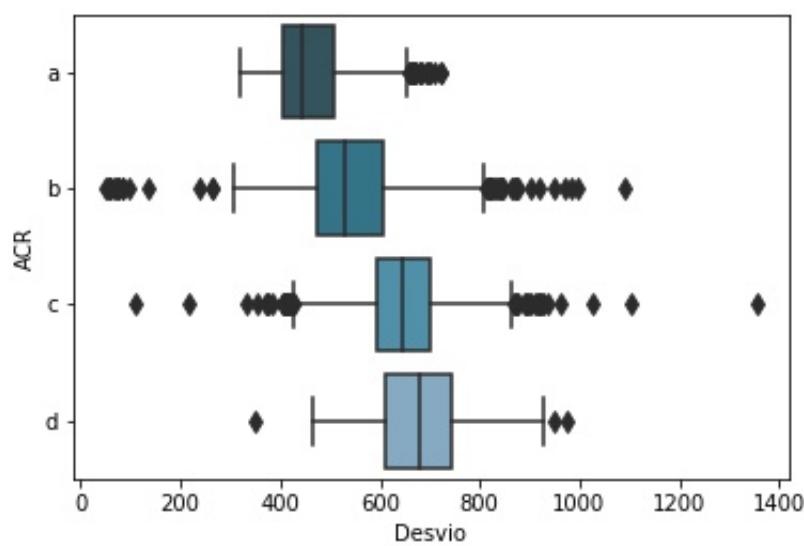


Figura 3.6.1: Variación del desvío estándar de los niveles de gris según la densidad mamaria.

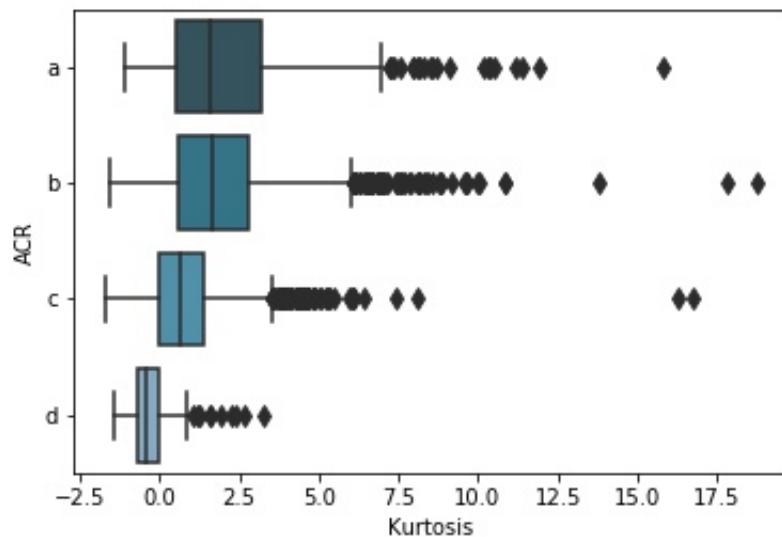


Figura 3.6.2: Variación de la curtosis de los niveles de gris según la densidad mamaria.

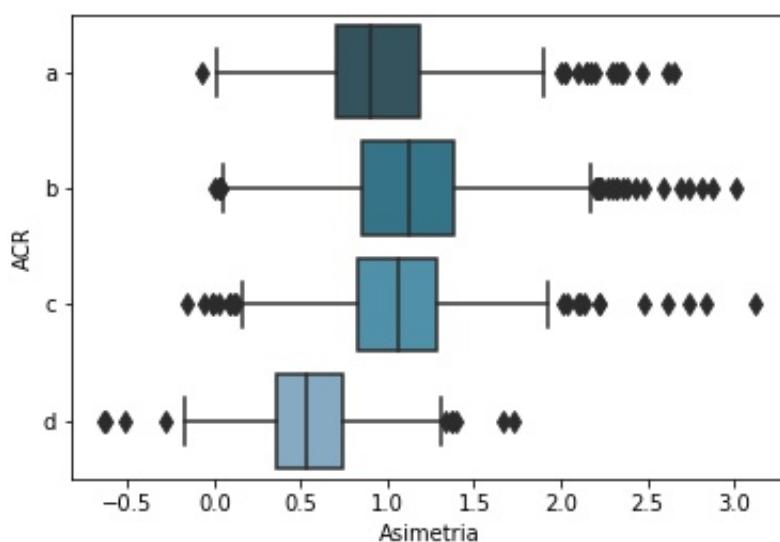


Figura 3.6.3: Variación de la asimetría de los niveles de gris según la densidad mamaria.

A este análisis se le sumó la variabilidad del tamaño del seno, proporcional a la cantidad de píxeles que forman la región de interés, con respecto a la densidad (ver en la Figura 3.6.4). Es posible observar que las mamas más pequeñas tienen mayor tendencia a tener una proporción de tejido fibroglandular superior.

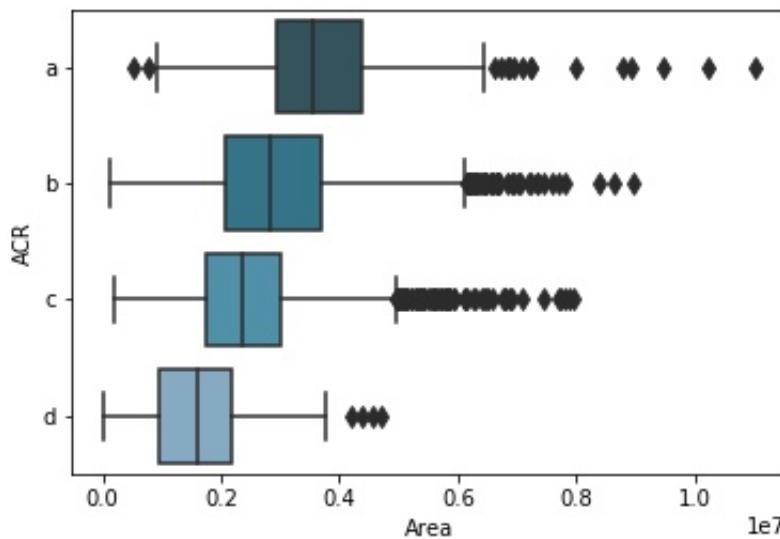


Figura 3.6.4: Relación entre el tamaño de la ROI y la densidad radiológica.

En la siguiente tabla se mostrará la descripción estadística de las características del histograma local a modo de resumen.

Tabla 3.6.1: Descripción de las características del histograma local

| | Area | Desvio_L1 | Desvio_L2 | Desvio_L3 | Asimetria_L1 | Asimetria_L2 | Asimetria_L3 |
|-------|-------------|-----------|-----------|-----------|--------------|--------------|--------------|
| count | 6396.00 | 6396.00 | 6396.00 | 6396.00 | 6396.00 | 6396.00 | 6396.00 |
| mean | 2723871.21 | 517.21 | 554.76 | 534.32 | 1.38 | 0.89 | 0.66 |
| std | 1265138.33 | 141.51 | 138.77 | 145.74 | 0.55 | 0.46 | 0.45 |
| min | 112201.00 | 52.93 | 49.49 | 0.00 | -0.62 | -1.17 | -1.56 |
| 25 % | 1858929.75 | 411.38 | 450.61 | 425.88 | 1.03 | 0.60 | 0.39 |
| 50 % | 2579437.50 | 503.54 | 551.30 | 528.57 | 1.34 | 0.87 | 0.67 |
| 75 % | 3461317.25 | 609.61 | 651.51 | 637.04 | 1.68 | 1.13 | 0.92 |
| max | 11013492.00 | 1203.05 | 1097.04 | 1609.01 | 5.26 | 5.35 | 3.34 |

2. Análisis de las características de la información DICOM

Ahora se analizarán los parámetros de adquisición de la imagen mamográfica y algunas otras características, obtenidos a partir de la información DICOM, de manera similar a cómo se hizo en el apartado anterior.

Para el caso de la edad (ver Figura 3.6.5), se pudo constatar que, en general, mujeres más jóvenes tienen senos más densos.

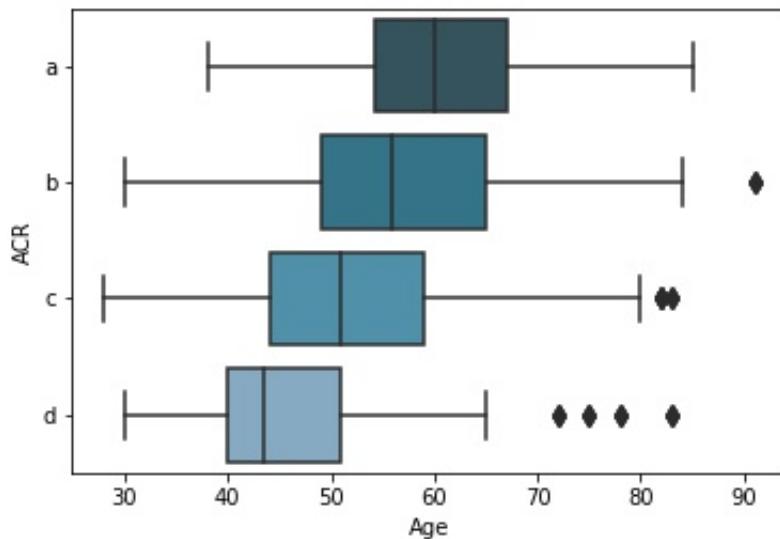


Figura 3.6.5: Variación de la densidad mamaria con la edad.

A partir del análisis de las Figuras 3.6.6, 3.6.9 y 3.6.8, es posible afirmar que el kilovoltaje pico, la fuerza de compresión y el espesor de la mama disminuyen con el aumento de la densidad

mamaria. Por otra parte, la corriente del tubo de rayos X (ver Figura 3.6.7), no tiene un aumento continuo como se esperaba, sino que aumenta hasta la clase “c” y luego disminuye.

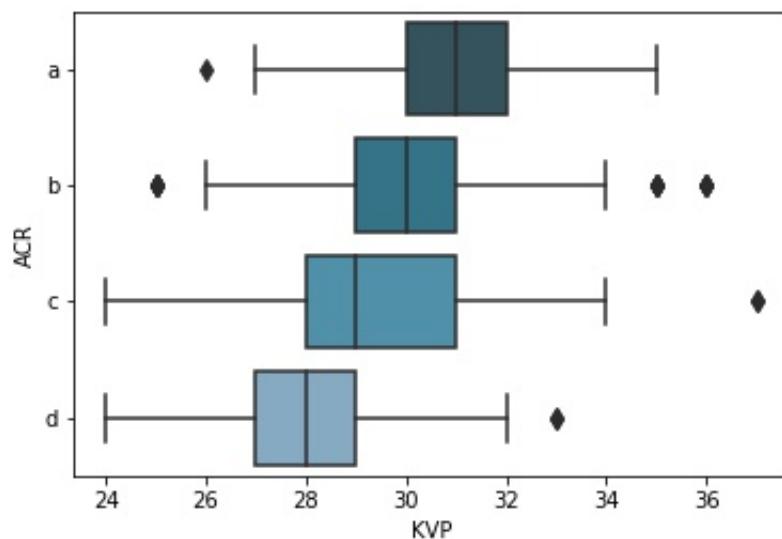


Figura 3.6.6: Relación entre el kilo voltaje pico utilizado y la densidad mamaria.

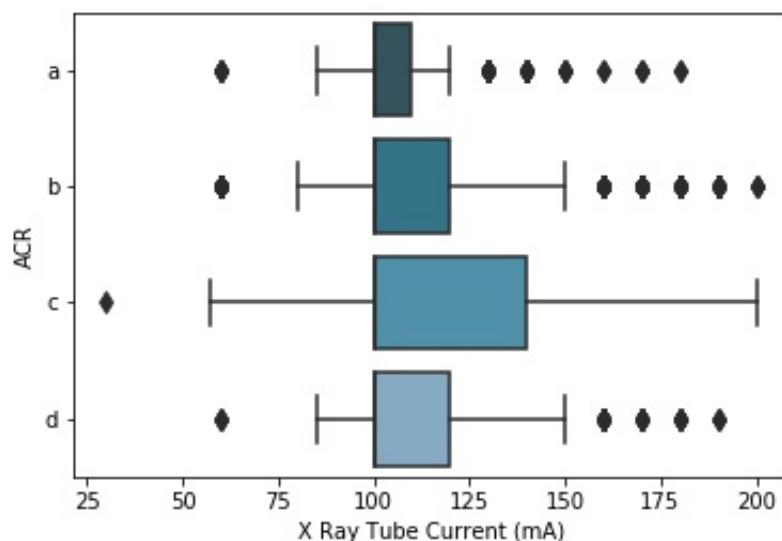


Figura 3.6.7: Relación entre la corriente (mA) del tubo de rayos X y la densidad mamaria.

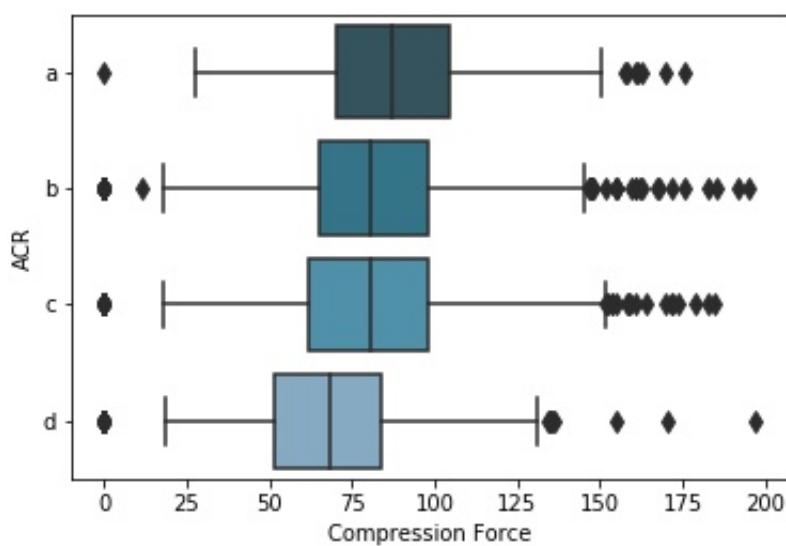


Figura 3.6.8: Relación entre la fuerza con la que se comprime la mama durante el estudio y la densidad mamaria.

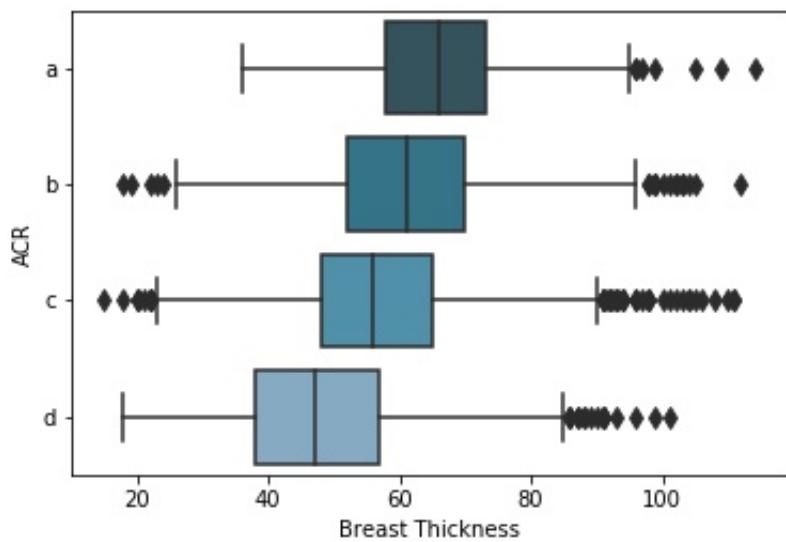


Figura 3.6.9: Relación entre el espesor de la mama comprimida durante el estudio y la densidad mamaria.

En la Tabla 3.6.3 se muestra una descripción resumida de la dispersión del conjunto de datos.

Tabla 3.6.2: Descripción de las características extraídas de la información DICOM.

| | Age | Dcm_1 | Dcm_2 | Dcm_3 | Dcm_4 | Dcm_6 | Dcm_9 | Dcm_10 | Dcm_12 |
|-------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| count | 6394.00 | 6394.00 | 6394.00 | 6394.00 | 6394.00 | 6394.00 | 6394.00 | 6394.00 | 6394.00 |
| mean | 54.59 | 29.69 | 116.22 | 127.29 | 1011.97 | 0.05 | 59.16 | 80.22 | 621.25 |
| std | 10.77 | 1.75 | 25.90 | 54.56 | 245.39 | 0.01 | 14.17 | 26.65 | 16.53 |
| min | 28.00 | 24.00 | 30.00 | 31.00 | 414.00 | 0.03 | 15.00 | 0.00 | 561.00 |
| 25 % | 46.00 | 28.00 | 100.00 | 85.00 | 819.00 | 0.05 | 50.00 | 63.00 | 609.00 |
| 50 % | 54.00 | 30.00 | 100.00 | 120.00 | 1116.00 | 0.05 | 59.00 | 80.00 | 621.00 |
| 75 % | 62.00 | 31.00 | 130.00 | 156.00 | 1168.00 | 0.05 | 68.00 | 97.89 | 640.00 |
| max | 91.00 | 37.00 | 200.00 | 446.00 | 2424.00 | 0.05 | 114.00 | 196.70 | 660.00 |

3. Análisis de características del gradiente

Seguiremos con el análisis de aquellas características que se han obtenido aplicando el operador gradiente a la imagen.

Se puede observar que ambas características, el promedio de los valores del módulo del gradiente a lo largo de la ROI (ver Figura 3.6.10) y la relación entre píxeles blancos sobre el total de píxeles de la misma en la imagen segmentada utilizando el operador gradiente, aumentan sutilmente con el aumento de la densidad mamaria.

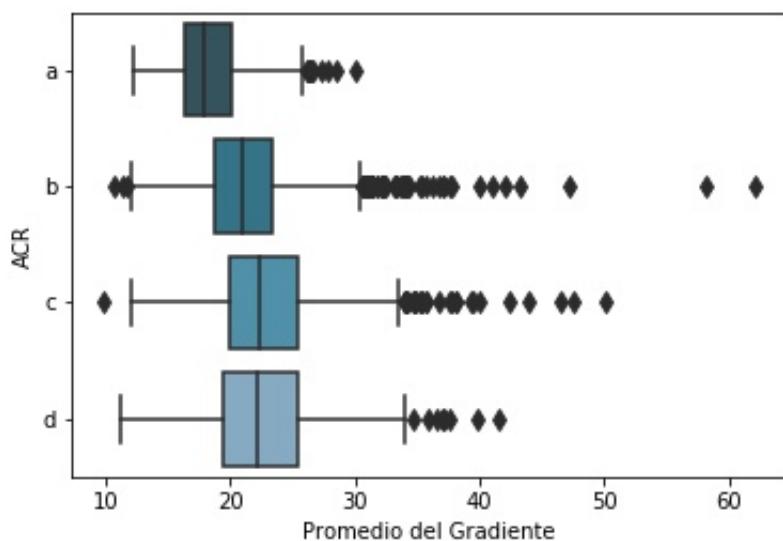


Figura 3.6.10: Relación entre el promedio del módulo del gradiente y la densidad mamaria.

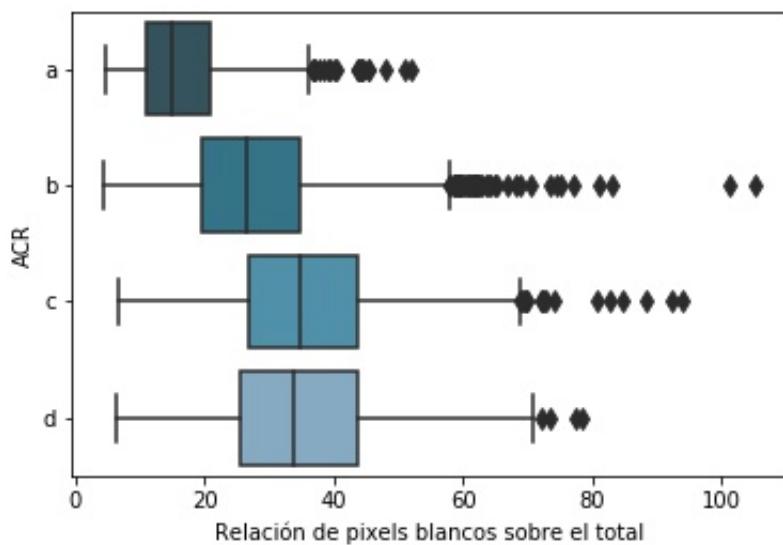


Figura 3.6.11: Diferencia en la relación de (pixels blancos/total de pixels) de la imagen con el filtro de gradiente aplicado con la densidad mamaria.

Tabla 3.6.3: Descripción de las características relacionadas con el operador gradiente.

| | promGradiente | Relacion | Relacion_L1 | Relacion_L2 | Relacion_L3 | Relacion_L4 |
|-------|---------------|----------|-------------|-------------|-------------|-------------|
| count | 6394.00 | 6394.00 | 6394.00 | 6394.00 | 6394.00 | 6394.00 |
| mean | 21.80 | 31.15 | 28.48 | 24.34 | 31.03 | 35.20 |
| std | 4.18 | 12.84 | 12.55 | 13.60 | 15.56 | 19.37 |
| min | 9.84 | 4.40 | 6.58 | 0.03 | 0.05 | 0.00 |
| 25 % | 18.91 | 21.76 | 19.42 | 13.89 | 19.06 | 20.26 |
| 50 % | 21.40 | 29.92 | 25.89 | 22.12 | 30.28 | 34.08 |
| 75 % | 24.27 | 39.55 | 34.78 | 32.96 | 42.00 | 49.12 |
| max | 62.03 | 105.19 | 97.11 | 94.71 | 82.70 | 100.00 |

3.6.2. Eliminación de datos faltantes

Los datos faltantes pueden implicar una reducción del tamaño de la muestra, lo cual nos puede impedir continuar con el análisis. Además, debemos asegurarnos de que el proceso de datos faltantes no esté sesgado y oculte una verdad inconveniente. Por lo tanto primero vemos cuáles son los datos faltantes en nuestro conjunto de datos:

| | Total | Percent |
|---------------------|-------|----------|
| RelEntropia | 145 | 0.021927 |
| EPBinaria | 145 | 0.021927 |
| EP | 145 | 0.021927 |
| Relacion_L4 | 65 | 0.009829 |
| Desvio_L3 | 65 | 0.009829 |
| Asimetria_L3 | 65 | 0.009829 |
| DFb1 | 25 | 0.003780 |
| DF1 | 25 | 0.003780 |
| DF5 | 22 | 0.003327 |
| DFb5 | 22 | 0.003327 |
| DFb6 | 22 | 0.003327 |
| Relacion_L3 | 22 | 0.003327 |
| DF6 | 22 | 0.003327 |
| Asimetria_L2 | 22 | 0.003327 |
| DF4 | 22 | 0.003327 |
| DF3 | 22 | 0.003327 |
| DFb4 | 22 | 0.003327 |
| Desvio_L2 | 22 | 0.003327 |
| DFb3 | 22 | 0.003327 |
| Asimetria_L1 | 21 | 0.003176 |

Figura 3.6.12: Atributos que contienen la mayor cantidad de datos faltantes.

Se probaron dos opciones:

- Reemplazar los datos faltantes por cero y por el promedio (con `sklearn.impute.SimpleImputer()`).
- Eliminar los datos faltantes (con `pandas.DataFrame.dropna ()`)

Se decidió eliminar los ejemplos que tenían datos faltantes, ya que se consideró que no era un número significativo frente al total de muestras, pasamos de tener seis mil seiscientos trece (6613) ejemplos a tener seis mil trescientos noventa y seis (6396).

3.6.3. Codificación de variables categóricas

Los datos categóricos son aquellos que toman sólo un número definido de valores. En la Figura 3.6.13 se muestran las variables de este tipo presentes en nuestra base de datos. Dado que, la mayoría de los algoritmos de aprendizaje requieren de datos continuos, es necesario codificarlos. Para ello hicimos uso del algoritmo One Hot Encoding (`pandas.get_dummies()`) que transforma cada característica categórica con n valores posibles en n características binarias, con solo una activa (ver Figura 3.6.14).

| | View | Laterality | Dcm_5 | Dcm_8 | Dcm_15 | Dcm_23 |
|-----------|-------------|-------------------|--------------|--------------|---------------|---------------|
| 0 | CC | R | RHODIUM | TUNGSTEN | AUTOMATIC | NO |
| 1 | MLO | R | RHODIUM | TUNGSTEN | AUTOMATIC | NO |
| 2 | MLO | L | RHODIUM | TUNGSTEN | AUTOMATIC | NO |
| 3 | CC | L | RHODIUM | TUNGSTEN | AUTOMATIC | NO |
| 4 | CC | R | SILVER | TUNGSTEN | AUTOMATIC | NO |
| 5 | MLO | R | SILVER | TUNGSTEN | AUTOMATIC | NO |
| 6 | MLO | L | SILVER | TUNGSTEN | AUTOMATIC | NO |
| 7 | CC | L | SILVER | TUNGSTEN | AUTOMATIC | NO |
| 8 | CC | L | MOLYBDENUM | MOLYBDENUM | AUTO_FILTER | NO |
| 9 | MLO | R | RHODIUM | MOLYBDENUM | AUTO_FILTER | NO |
| 10 | CC | R | MOLYBDENUM | MOLYBDENUM | AUTO_FILTER | NO |
| 11 | MLO | L | MOLYBDENUM | MOLYBDENUM | AUTO_FILTER | NO |
| 12 | CC | R | RHODIUM | TUNGSTEN | AUTOMATIC | NO |
| 13 | MLO | R | RHODIUM | TUNGSTEN | AUTOMATIC | NO |
| 14 | MLO | L | RHODIUM | TUNGSTEN | AUTOMATIC | NO |
| 15 | CC | L | RHODIUM | TUNGSTEN | AUTOMATIC | NO |
| 16 | CC | R | RHODIUM | TUNGSTEN | AUTOMATIC | NO |
| 17 | CC | L | SILVER | TUNGSTEN | AUTOMATIC | NO |
| 18 | MLO | R | SILVER | TUNGSTEN | AUTOMATIC | NO |
| 19 | MLO | L | SILVER | TUNGSTEN | AUTOMATIC | NO |

Figura 3.6.13: Variables categóricas.

| | View_CC | View_MLO | Laterality_L | Dcm_5_MOLYBDENUM | Dcm_5_RHODIUM | Dcm_5_SILVER |
|----|----------------|-----------------|---------------------|-------------------------|----------------------|---------------------|
| 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| 2 | 0 | 1 | 1 | 0 | 1 | 0 |
| 3 | 1 | 0 | 1 | 0 | 1 | 0 |
| 4 | 1 | 0 | 0 | 0 | 0 | 1 |
| 5 | 0 | 1 | 0 | 0 | 0 | 1 |
| 6 | 0 | 1 | 1 | 0 | 0 | 1 |
| 7 | 1 | 0 | 1 | 0 | 0 | 1 |
| 8 | 1 | 0 | 1 | 1 | 0 | 0 |
| 9 | 0 | 1 | 0 | 0 | 1 | 0 |
| 10 | 1 | 0 | 0 | 1 | 0 | 0 |
| 11 | 0 | 1 | 1 | 1 | 0 | 0 |
| 12 | 1 | 0 | 0 | 0 | 1 | 0 |
| 13 | 0 | 1 | 0 | 0 | 1 | 0 |
| 14 | 0 | 1 | 1 | 0 | 1 | 0 |
| 15 | 1 | 0 | 1 | 0 | 1 | 0 |
| 16 | 1 | 0 | 0 | 0 | 1 | 0 |
| 17 | 1 | 0 | 1 | 0 | 0 | 1 |
| 18 | 0 | 1 | 0 | 0 | 0 | 1 |
| 19 | 0 | 1 | 1 | 0 | 0 | 1 |

Figura 3.6.14: Ejemplos de variables categóricas codificadas.

3.6.4. Escalamiento de los datos

La estandarización de los conjuntos de datos es un requisito común para muchos estimadores de aprendizaje automático implementados en la librería de Python scikit-learn; podrían comportarse mal si las características individuales no se parecen más o menos a los datos estándar distribuidos normalmente: gaussianos con media cero y variación de unidades.

Para el escalamiento de los datos se probaron las siguientes opciones:

- **sklearn.preprocessing.StandardScaler()**: Estandariza las características mediante la eliminación de la media y las escala a la varianza unitaria. El centrado y la escalación se realizan de forma independiente en cada función al calcular las estadísticas relevantes de las muestras en el conjunto de entrenamiento. La media y la desviación estándar se almacenan para ser utilizadas en datos posteriores utilizando el método de transformación.
- **sklearn.preprocessing.MinMaxScaler()**: Transforma las características al escalar cada una de ellas a un rango dado. Por defecto entre 0 y 1.
- **sklearn.preprocessing.RobustScaler()**: Este escalador elimina la mediana y escala los datos de acuerdo con el rango de cuantiles (por defecto, IQR: Interquartile Range). El IQR es el rango entre el primer cuartil (cuantil 25) y el tercer cuartil (cuantil 75). El centrado y la escalamiento

se realizan de forma independiente en cada función al calcular las estadísticas relevantes de las muestras en el conjunto de entrenamiento.

Con el que se obtuvo un mejor desempeño de los clasificadores fue con el StandardScaler.

3.6.5. Reducción de dimensionalidad

Se aplicó el algoritmo de ACP (Análisis de Componentes Principales) para reducir la dimensionalidad de nuestra base de datos y así disminuir tiempos de cómputo.

Se obtuvieron 30 componentes ortogonales entre sí que explican el 95 % de la variabilidad de los datos.

3.7. Ajuste de los estimadores

Entrenar modelos de aprendizaje automático requiere dos tipos de parámetros: los que se aprenden de los datos y los del algoritmo. Estos últimos son los parámetros de ajuste, también llamados hiperparámetros, por ejemplo C , kernel y gamma para Support Vector Classifier.

Muchos de los modelos que se implementan en aprendizaje automático requieren que se fijen los valores de los hiperparámetros durante el entrenamiento. En la mayoría de las ocasiones, la selección de un valor u otro no es una tarea trivial. Se debe tener en cuenta que es común que un subconjunto de esos parámetros tenga un gran impacto en el rendimiento predictivo o de cálculo del modelo, mientras que otros pueden dejarse a sus valores predeterminados. Es posible seleccionar los parámetros más apropiados para un modelo y un conjunto de datos utilizando la técnica de validación cruzada. Para este propósito se puede utilizar una herramienta disponible en scikit-learn, llamada **GridSearchCV**.

3.7.1. Validación cruzada

La validación cruzada (CV) es una técnica con la que se puede identificar la existencia de diferentes problemas durante el entrenamiento de los modelos, como la aparición de sobreajuste. Permitiendo así obtener modelos más estables.

En la validación cruzada de K iteraciones o K-fold cross-validation los datos de muestra se dividen en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. El proceso de validación cruzada es repetido durante k iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. Finalmente se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado. Este método es muy preciso puesto que evaluamos a partir de K combinaciones de datos de entrenamiento y de prueba, pero aun así tiene una desventaja, y es que, a diferencia del método de retención, es lento desde el punto de vista computacional. En la práctica, la elección del número de iteraciones depende de la medida del conjunto de datos. Para este trabajo se eligió k= 10 (ver Figura 3.7.1).

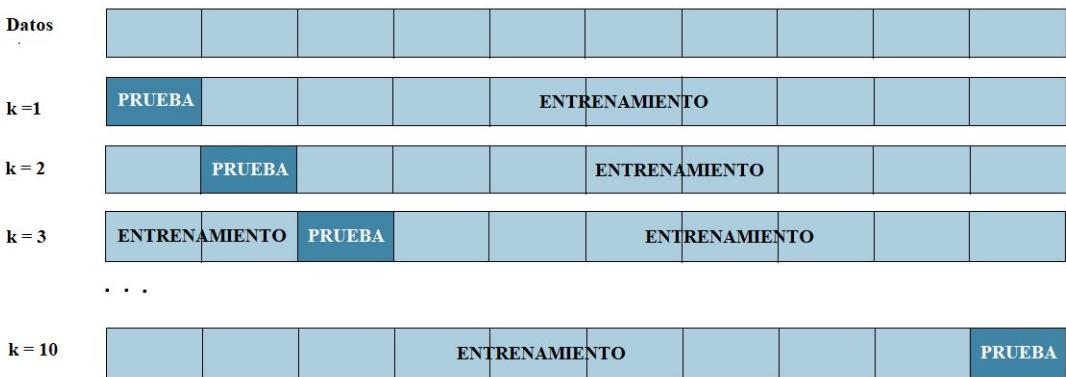


Figura 3.7.1: Validación cruzada.

3.7.2. Selección de parámetros utilizando GridSearchCV

La búsqueda en cuadrícula proporcionada por GridSearchCV genera de manera exhaustiva candidatos de una grilla de valores. Esta clase se ha de llamar indicándole la instancia de un modelo, los valores a probar y el número de conjuntos en el que se dividen los datos. Esto se realiza mediante los siguientes parámetros:

- **estimator**: el modelo que se ha de evaluar (SVC, Random Forest, k-NN, Regresión Logística).
- **param_grid**: Diccionario con nombres de parámetros (cadena) como claves y listas de configuraciones de parámetros para probar como valores, o una lista de dichos diccionarios, en cuyo caso se exploran las cuadrículas abarcadas por cada diccionario en la lista. Esto permite buscar en cualquier secuencia de configuraciones de parámetros.
- **cv**: el número de conjuntos en los que se divide los datos para la validación cruzada.

Se obtienen como resultado:

- El estimador que fue elegido por la búsqueda, es decir, el estimador que dio el puntaje más alto (**best_estimator_**).
- Puntuación media validada cruzada del mejor estimador (**best_score_**).
- Configuración de parámetros que dio los mejores resultados en los datos de espera (**best_params_**).

Al evaluar el modelo resultante, es importante hacerlo en muestras retenidas que no se vieron durante el proceso de búsqueda de grillas. Por este motivo se dividieron los datos en un conjunto de desarrollo (para ser alimentado a la instancia de GridSearchCV) y un conjunto de evaluación para calcular las métricas de rendimiento. Lo cual se realizó utilizando la función de **train_test_split**.

Capítulo 4

Resultados

4.1. Clasificación con la base de datos segmentada

Con el fin de evaluar el desempeño en la clasificación de la base de datos, se decidió agrupar los atributos según criterios de similitud:

- Características extraídas de histogramas.
- Características extraídas de la información DICOM.
- Características extraídas de la imagen con los valores del módulo del gradiente.
- Características relacionadas a la entropía de permutaciones.
- Descriptores de textura de Haralick.
- Características relacionadas a la dimensión fractal de la imagen.
- Características relacionadas a la dimensión fractal de los bordes de la imagen.

Posteriormente se aplicaron dos clasificadores: SVC (Support Vector Classifier, de la familia de SVM) Y k-NN(K Nearest Neighbors), a cada conjunto por separado. Como los resultados obtenidos con ambos fueron similares, se expondrán solamente los alcanzados con k-NN.

En los reportes de clasificación se pueden analizar los siguientes elementos para cada una de las clases:

- *precision*: es la relación $V_p/(V_p + F_p)$ donde V_p es el número de verdaderos positivos y F_p el número de falsos positivos. La precisión es intuitivamente la capacidad del clasificador para no etiquetar como positiva una muestra que es negativa.
- *recall*: es la relación $V_p/(V_p + F_n)$ donde V_p es el número de verdaderos positivos y F_n el número de falsos negativos. Recall es intuitivamente la capacidad del clasificador para encontrar todas las muestras positivas.

- *f1-score*: puede interpretarse como una media armónica ponderada de la precisión y el recall, donde una puntuación F-beta alcanza su mejor valor en 1 y la peor puntuación en 0. Pondera el recall más que la precisión por un factor de beta. beta = 1.0 significa recuperación y precisión son igualmente importantes.
- *support*: es el número de ocurrencias de cada clase en *y_test*.

Por otra parte, se tiene la *matriz de confusión*, que se utiliza para evaluar la calidad de la salida de un clasificador. En el eje horizontal se tienen a las clases predichas, mientras que en el vertical se tienen a las clases verdaderas, de este modo, los elementos sobre la diagonal representan el número de puntos para los cuales la clase predicha es igual a la clase verdadera, mientras que los elementos fuera de la diagonal son aquellos que están clasificados incorrectamente por el clasificador. Cuanto más altos sean los valores sobre la diagonal de la matriz de confusión, mejor, ya que indica muchas predicciones correctas.

| | precision | recall | f1-score | support |
|-------------|-----------|--------|----------|---------|
| a | 0.63 | 0.25 | 0.36 | 95 |
| b | 0.71 | 0.81 | 0.76 | 518 |
| c | 0.77 | 0.79 | 0.78 | 570 |
| d | 0.70 | 0.44 | 0.54 | 97 |
| avg / total | 0.73 | 0.73 | 0.72 | 1280 |

Figura 4.1.1: Reporte de la clasificación utilizando características extraídas del histograma y k-NN.

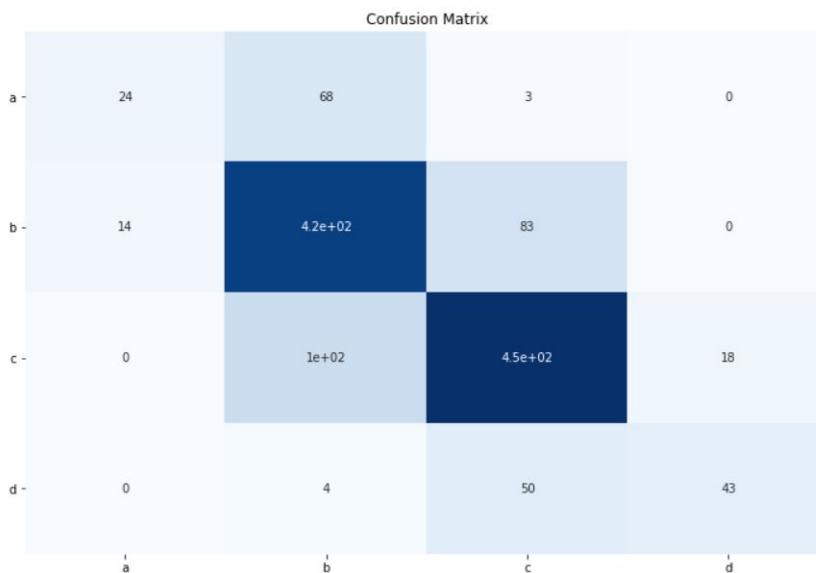


Figura 4.1.2: Matriz de confusión.

| | precision | recall | f1-score | support |
|-------------|-----------|--------|----------|---------|
| a | 1.00 | 0.01 | 0.02 | 103 |
| b | 0.60 | 0.75 | 0.67 | 510 |
| c | 0.69 | 0.75 | 0.72 | 572 |
| d | 0.61 | 0.15 | 0.24 | 94 |
| avg / total | 0.68 | 0.65 | 0.61 | 1279 |

Figura 4.1.3: Reporte de la clasificación utilizando características extraídas de la información DICOM y k-NN.

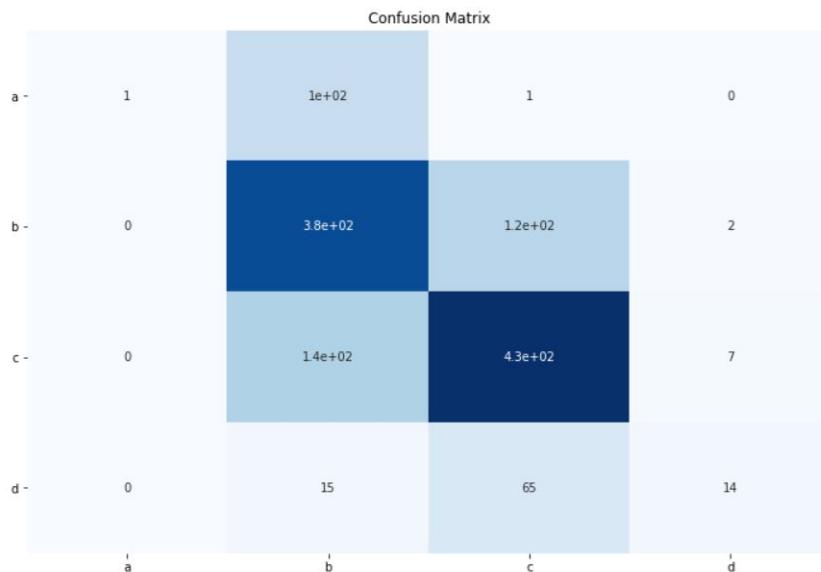


Figura 4.1.4: Matriz de confusión.

| | precision | recall | f1-score | support |
|-------------|-----------|--------|----------|---------|
| a | 0.70 | 0.36 | 0.47 | 103 |
| b | 0.65 | 0.73 | 0.69 | 510 |
| c | 0.70 | 0.77 | 0.73 | 572 |
| d | 0.45 | 0.10 | 0.16 | 94 |
| avg / total | 0.66 | 0.67 | 0.65 | 1279 |

Figura 4.1.5: Reporte de la clasificación utilizando características extraídas de la imagen a la que se le aplicó el operador gradiente y k-NN.

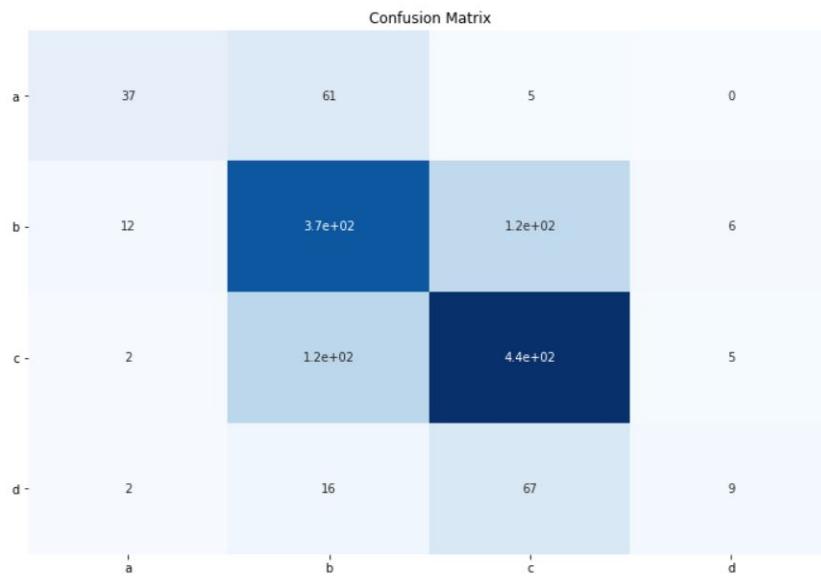


Figura 4.1.6: Matriz de confusión.

| | precision | recall | f1-score | support |
|-------------|-----------|--------|----------|---------|
| a | 0.67 | 0.08 | 0.14 | 103 |
| b | 0.66 | 0.77 | 0.71 | 510 |
| c | 0.73 | 0.79 | 0.76 | 572 |
| d | 0.72 | 0.33 | 0.45 | 94 |
| avg / total | 0.69 | 0.69 | 0.67 | 1279 |

Figura 4.1.7: Reporte de la clasificación utilizando los descriptores de textura de Haralick de la imagen y k-NN.

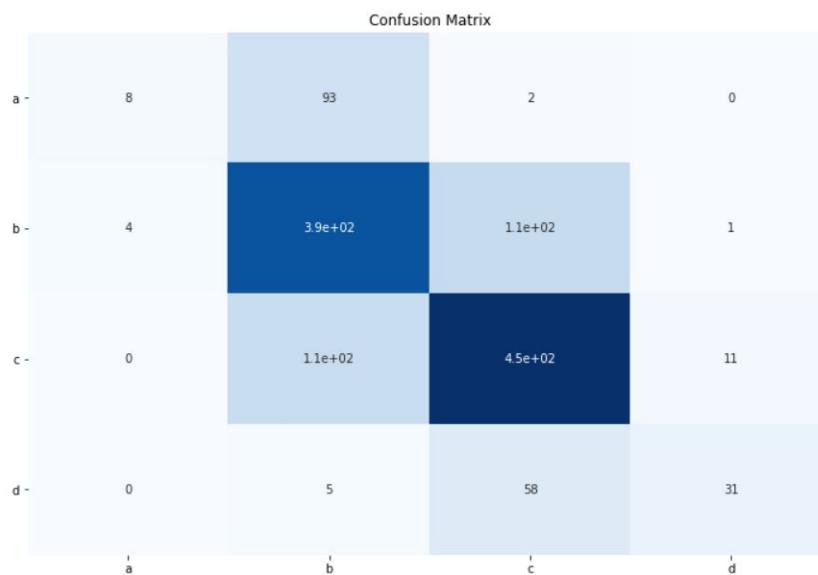


Figura 4.1.8: Matriz de confusión.

| | precision | recall | f1-score | support |
|-------------|-----------|--------|----------|---------|
| a | 0.53 | 0.19 | 0.28 | 103 |
| b | 0.57 | 0.57 | 0.57 | 510 |
| c | 0.59 | 0.76 | 0.67 | 572 |
| d | 0.00 | 0.00 | 0.00 | 94 |
| avg / total | 0.53 | 0.58 | 0.55 | 1279 |

Figura 4.1.9: Reporte de la clasificación utilizando las características relacionadas con la entropía de permutaciones y k-NN.

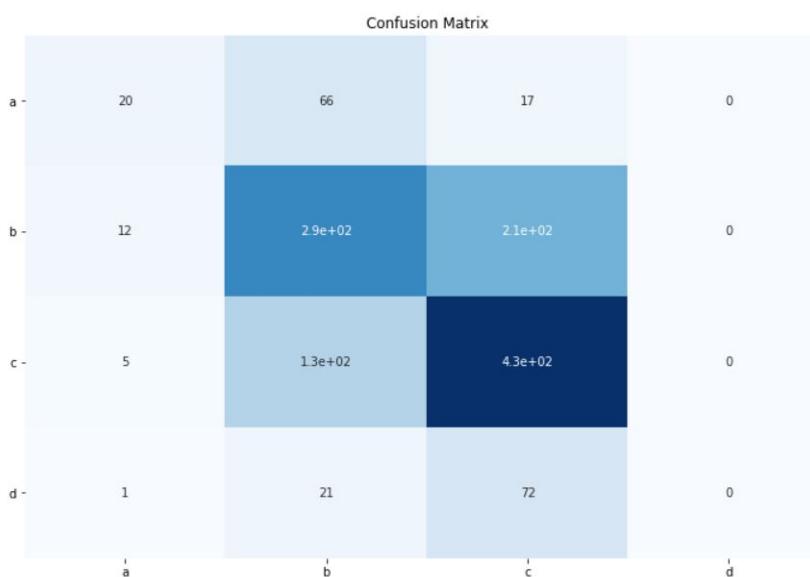


Figura 4.1.10: Matriz de confusión.

| | precision | recall | f1-score | support |
|-------------|-----------|--------|----------|---------|
| a | 0.59 | 0.25 | 0.35 | 103 |
| b | 0.69 | 0.78 | 0.73 | 510 |
| c | 0.76 | 0.80 | 0.78 | 572 |
| d | 0.71 | 0.41 | 0.52 | 94 |
| avg / total | 0.71 | 0.72 | 0.71 | 1279 |

Figura 4.1.11: Reporte de la clasificación utilizando las características relacionadas con la dimensión fractal de la imagen y k-NN.

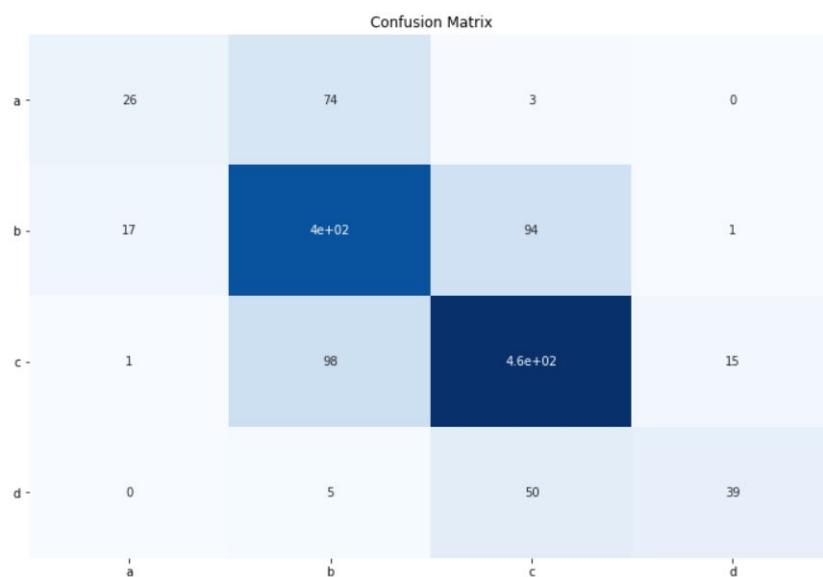


Figura 4.1.12: Matriz de confusión.

| | precision | recall | f1-score | support |
|-------------|-----------|--------|----------|---------|
| a | 0.55 | 0.27 | 0.36 | 103 |
| b | 0.68 | 0.79 | 0.73 | 510 |
| c | 0.77 | 0.78 | 0.77 | 572 |
| d | 0.70 | 0.45 | 0.55 | 94 |
| avg / total | 0.71 | 0.72 | 0.71 | 1279 |

Figura 4.1.13: Reporte de la clasificación utilizando las características relacionadas con la dimensión fractal de los bordes de la de la imagen y k-NN.

| | | Confusion Matrix | | | |
|-----|----|------------------|---------|-----|-----|
| | | a - | b - | c - | d - |
| a - | 28 | 73 | 2 | 0 | |
| | 22 | 4e+02 | 86 | 0 | |
| c - | 1 | 1.1e+02 | 4.4e+02 | 18 | |
| | 0 | 8 | 44 | 42 | |
| | | a | b | c | d |

Figura 4.1.14: Matriz de confusión.

A partir de las definiciones dadas y el estudio de los de los reportes y las matrices de confusión, se pueden destacar tanto las características extraídas del histograma (ver Figura 4.1.1) como las relacionadas con la dimensión fractal (ver Figura 4.1.11 y 4.1.13), las mismas demostraron que la información que aportan en la descripción de la imagen es suficiente para obtener una buena clasificación de los datos.

Además, a excepción de la entropía de permutaciones (ver Figura 4.1.9), todos los descriptores de textura dieron resultados satisfactorios. Éstos son: características extraídas de la imagen de los valores del módulo de del gradiente, descriptores de Haralick y dimensión fractal. De esta manera, podemos afirmar que se cumplió uno de nuestros objetivos: demostrar que existe una relación entre la textura de la imagen y la densidad radiológica de la misma.

4.2. Clasificación con la base de datos completa

Se probaron cuatro diferentes algoritmos inteligentes de agrupamiento para poder determinar cuál entrega el mejor resultado. A su vez, para cada uno de estos estimadores se realizó una optimización de los hiperparámetros (configuraciones del modelo). Esto significa encontrar la combinación de valores de los hiperparámetros que permita obtener la mejor puntuación de validación cruzada para un modelo de aprendizaje determinado.

Desempeño del algoritmo de Random Forest

Al reporte de clasificación descripto en la sección anterior se le suman otro tres elementos:

- *micro average*: micro promedios. Promedio del total de positivos verdaderos, falsos negativos y falsos positivos.

- *macro average*: promedio macro. Promedio de la media no ponderada por etiqueta.
- *weighted average*: promedio ponderado. Promedio de la media ponderada por soporte por etiqueta.

También se agregó la matriz de confusión con los valores de los errores.

Analizando las Figuras 4.2.1 a 4.2.12 se puede advertir que el valor de recall es mucho menor para las clases “a” y “d”, esto hace que la proporción de imágenes clasificadas incorrectamente aumente para dichos casos. Creemos que esto puede deberse a la diferencia en la cantidad de ejemplos, es por ello que se propone como trabajo a futuro igualar dichas cantidades y evaluar si los promedios de acierto mejoran.

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| a | 0.71 | 0.41 | 0.52 | 95 |
| b | 0.72 | 0.81 | 0.76 | 518 |
| c | 0.77 | 0.79 | 0.78 | 570 |
| d | 0.74 | 0.47 | 0.58 | 97 |
| micro avg | 0.75 | 0.75 | 0.75 | 1280 |
| macro avg | 0.74 | 0.62 | 0.66 | 1280 |
| weighted avg | 0.75 | 0.75 | 0.74 | 1280 |

Figura 4.2.1: Reporte de la clasificación utilizando Random Forest.

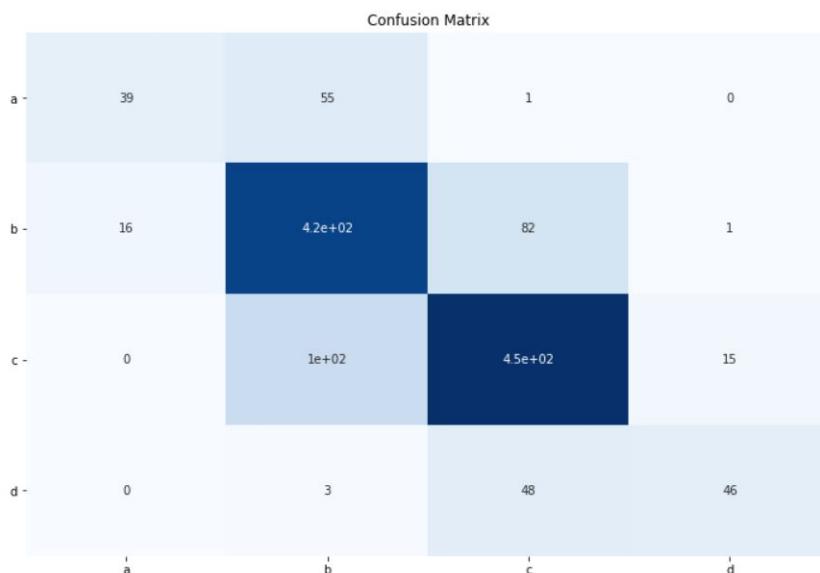


Figura 4.2.2: Matriz de confusión para la clasificación con Random Forest.

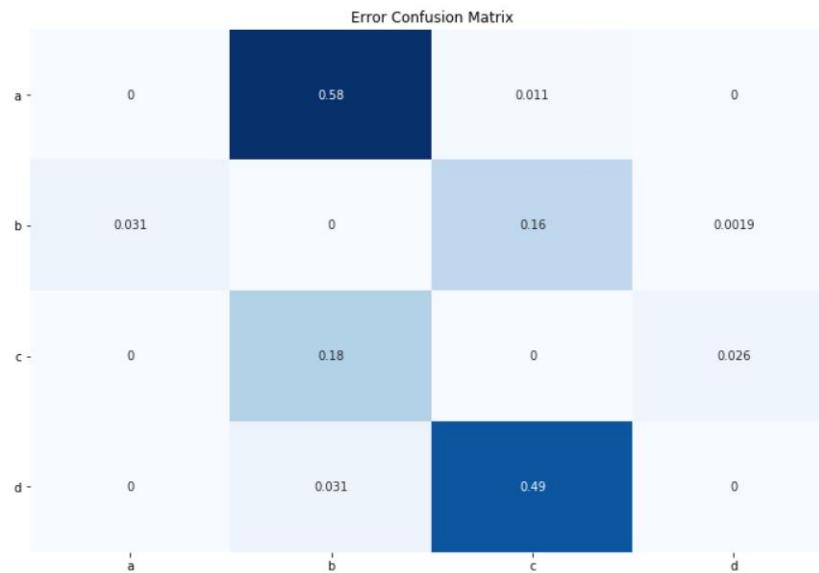


Figura 4.2.3: Matriz de confusión de los errores para la clasificación con Random Forest.

Desempeño del algoritmo de Regresión Logística

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| a | 0.62 | 0.34 | 0.44 | 95 |
| b | 0.73 | 0.83 | 0.77 | 518 |
| c | 0.80 | 0.79 | 0.80 | 570 |
| d | 0.68 | 0.58 | 0.63 | 97 |
| micro avg | 0.75 | 0.75 | 0.75 | 1280 |
| macro avg | 0.71 | 0.63 | 0.66 | 1280 |
| weighted avg | 0.75 | 0.75 | 0.75 | 1280 |

Figura 4.2.4: Reporte de la clasificación utilizando Regresión Logística.

| Error Confusion Matrix Linear Regression | | | | |
|--|-------|------|------|--------|
| a - | 0 | 0.66 | 0 | 0 |
| b - | 0.039 | 0 | 0.13 | 0.0039 |
| c - | 0 | 0.17 | 0 | 0.042 |
| d - | 0 | 0 | 0.42 | 0 |
| | 'a' | 'b' | 'c' | 'd' |

Figura 4.2.5: Matriz de confusión para la clasificación con Regresión Logística.

| Error Confusion Matrix Linear Regression | | | | |
|--|-------|------|------|--------|
| a - | 0 | 0.66 | 0 | 0 |
| b - | 0.039 | 0 | 0.13 | 0.0039 |
| c - | 0 | 0.17 | 0 | 0.042 |
| d - | 0 | 0 | 0.42 | 0 |
| | 'a' | 'b' | 'c' | 'd' |

Figura 4.2.6: Matriz de confusión de los errores para la clasificación con Regresión Logística.

Desempeño del algoritmo de k-Vecinos

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| a | 0.65 | 0.18 | 0.28 | 95 |
| b | 0.69 | 0.81 | 0.74 | 518 |
| c | 0.75 | 0.79 | 0.77 | 570 |
| d | 0.78 | 0.33 | 0.46 | 97 |
| micro avg | 0.72 | 0.72 | 0.72 | 1280 |
| macro avg | 0.72 | 0.53 | 0.57 | 1280 |
| weighted avg | 0.72 | 0.72 | 0.70 | 1280 |

Figura 4.2.7: Reporte de la clasificación utilizando k-NN.

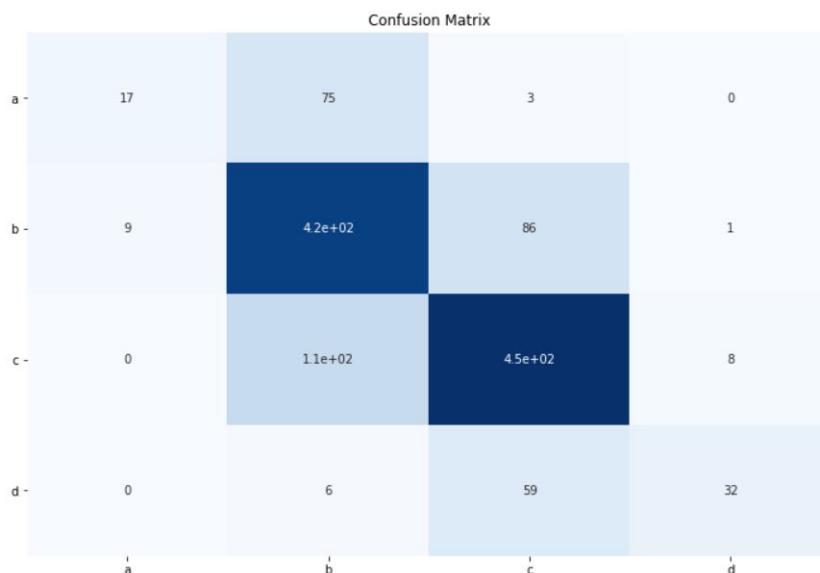


Figura 4.2.8: Matriz de confusión para la clasificación con k-NN.

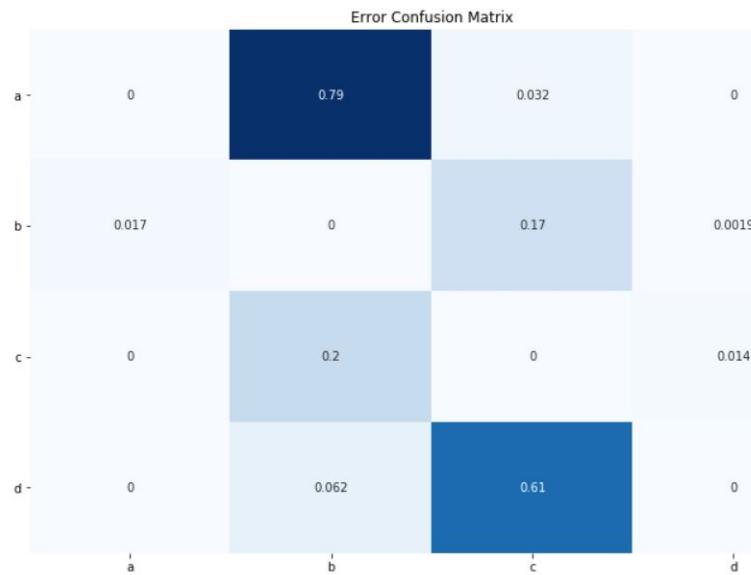


Figura 4.2.9: Matriz de confusión de los errores para la clasificación con k-NN.

Desempeño del algoritmo de SVC

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| a | 0.73 | 0.45 | 0.56 | 95 |
| b | 0.73 | 0.83 | 0.78 | 518 |
| c | 0.81 | 0.79 | 0.80 | 570 |
| d | 0.74 | 0.60 | 0.66 | 97 |
| micro avg | 0.76 | 0.76 | 0.76 | 1280 |
| macro avg | 0.75 | 0.67 | 0.70 | 1280 |
| weighted avg | 0.77 | 0.76 | 0.76 | 1280 |

Figura 4.2.10: Reporte de la clasificación utilizando SVC.

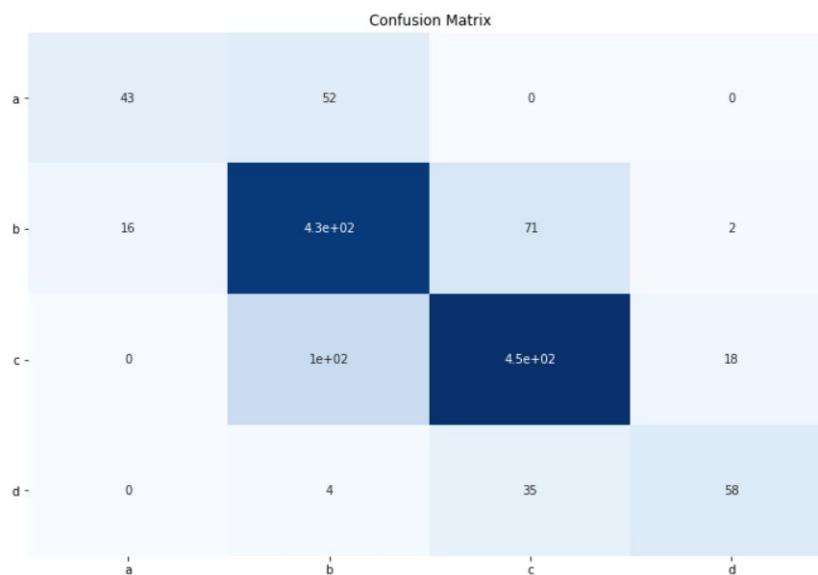


Figura 4.2.11: Matriz de confusión para la clasificación con SVC.

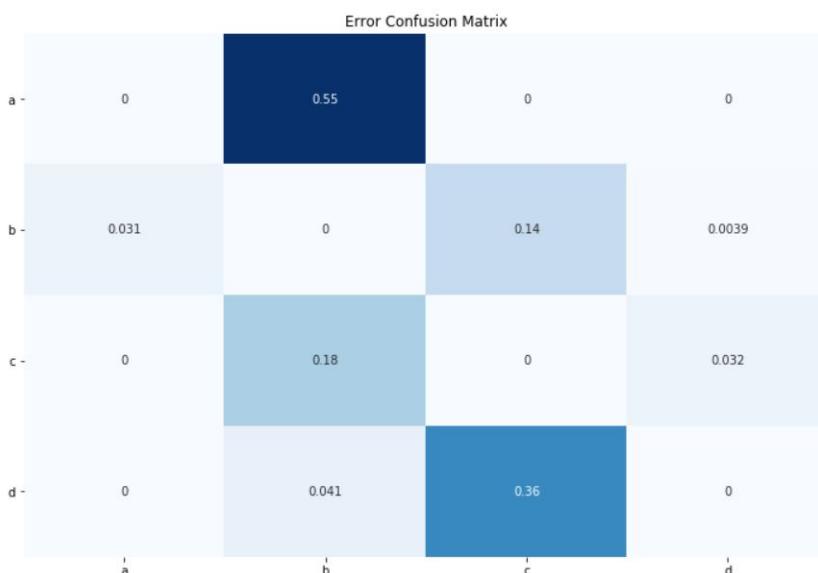


Figura 4.2.12: Matriz de confusión de los errores para la clasificación con SVC.

La Tabla 4.2.1 resume los resultados entregados por cada uno de los de los clasificadores. Ha de notarse que el mejor promedio de aciertos (accuracy) se obtuvo con SVC.

Tabla 4.2.1: Comparación de los resultados obtenidos con los diferentes algoritmos.

| Score | SVC | LR | RF | k-NN |
|-------------------------|-------|-------|-------|-------|
| 0 Test Accuracy | 0.764 | 0.753 | 0.746 | 0.720 |
| 1 Mean Cross Validation | 0.754 | 0.747 | 0.722 | 0.717 |

4.3. Clasificación luego de aplicar ACP

Finalmente, se realizó una reducción de las dimensiones de la base de datos y nuevamente se experimentó con diferentes algoritmos de clasificación (ver Figura 4.3.1) para evaluar su desempeño y quedarnos con el mejor modelo.

```
Performing model optimizations...

Estimator: AdaBoost
Best params: {'clf_n_estimators': 100}
Best training accuracy: 0.607
Test set accuracy score for best params: 0.607

Estimator: Logistic Regression
Best params: {'clf_C': 0.5, 'clf_penalty': 'l2', 'clf_solver': 'liblinear'}
Best training accuracy: 0.733
Test set accuracy score for best params: 0.737

Estimator: K-Nearest Neighbor
Best params: {'clf_n_neighbors': 20}
Best training accuracy: 0.722
Test set accuracy score for best params: 0.710

Estimator: Random Forest
Best params: {'clf_criterion': 'entropy', 'clf_max_depth': 10, 'clf_min_samples_leaf': 2, 'clf_min_samples_split': 2, 'clf_n_estimators': 100}
Best training accuracy: 0.737
Test set accuracy score for best params: 0.741

Estimator: Extra Trees
Best params: {'clf_criterion': 'entropy', 'clf_max_depth': 10, 'clf_min_samples_leaf': 2, 'clf_min_samples_split': 10, 'clf_n_estimators': 100}
Best training accuracy: 0.718
Test set accuracy score for best params: 0.710

Estimator: Support Vector Machine
Best params: {'clf_C': 2, 'clf_kernel': 'rbf'}
Best training accuracy: 0.754
Test set accuracy score for best params: 0.748

Classifier with best test set accuracy: Support Vector Machine
```

Figura 4.3.1: Experimentación con seis diferentes algoritmos.

Se pudo constatar que el mejor modelo está dado por el clasificador SVC, que si bien el promedio de aciertos es sutilmente menor, se considera insignificante frente a la reducción de tiempos de cómputo.

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| a | 0.69 | 0.39 | 0.50 | 95 |
| b | 0.70 | 0.82 | 0.76 | 518 |
| c | 0.80 | 0.77 | 0.79 | 570 |
| d | 0.79 | 0.57 | 0.66 | 97 |
| micro avg | 0.75 | 0.75 | 0.75 | 1280 |
| macro avg | 0.74 | 0.64 | 0.68 | 1280 |
| weighted avg | 0.75 | 0.75 | 0.74 | 1280 |

Figura 4.3.2: Reporte de la clasificación utilizando ACP y SVC.

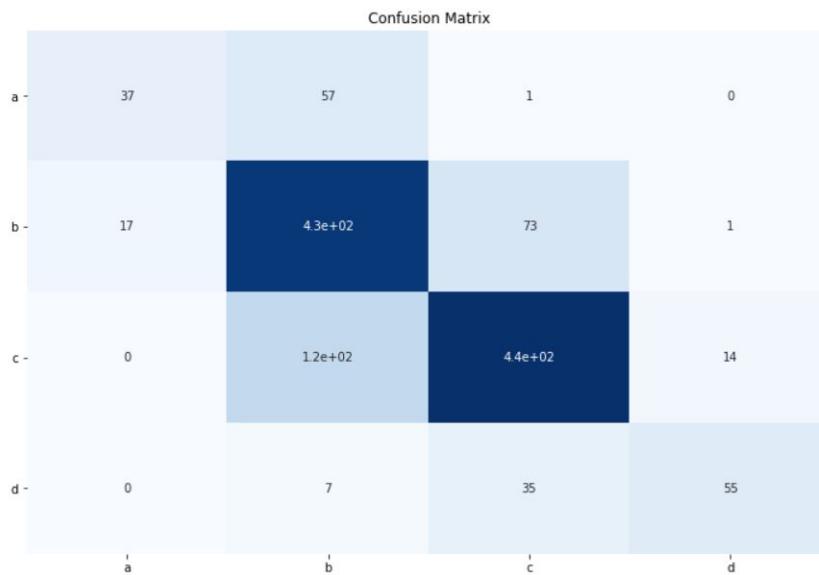


Figura 4.3.3: Matriz de confusión para la clasificación con ACP y SVC.

4.4. Tratamiento del desbalance entre clases

La diferencia entre las cantidades de ejemplos por clase afecta la cantidad de aciertos del modelo elegido. Para intentar saldar este problema, se decidió triplicar el número de ejemplos de las clases “a” y “d”, y evaluar una vez más, el desempeño de los clasificadores.

En este caso se eligieron SVC Y Random Forest, obteniendo los siguientes resultados:

Tabla 4.4.1: Comparación de los resultados obtenidos luego de triplicar los ejemplos de las clases “a” y “d”.

| Score | SVC | Random Forest |
|-------------------------|-------|---------------|
| 0 Test Accuracy | 0.759 | 0.752 |
| 1 Mean Cross Validation | 0.770 | 0.825 |

Es posible observar en la Tabla 4.4.1 que si bien el valor en la validación cruzada (durante el entrenamiento del modelo), aumenta considerablemente, particularmente para Random Forest, no es así durante la etapa de prueba. Aunque se debe destacar que los errores se distribuyen de manera más uniforme entre las clases:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| a | 0.67 | 0.72 | 0.69 | 132 |
| b | 0.78 | 0.72 | 0.75 | 539 |
| c | 0.78 | 0.81 | 0.80 | 552 |
| d | 0.64 | 0.73 | 0.68 | 86 |
| micro avg | 0.76 | 0.76 | 0.76 | 1309 |
| macro avg | 0.72 | 0.75 | 0.73 | 1309 |
| weighted avg | 0.76 | 0.76 | 0.76 | 1309 |

Figura 4.4.1: Reporte de la clasificación con SVC.

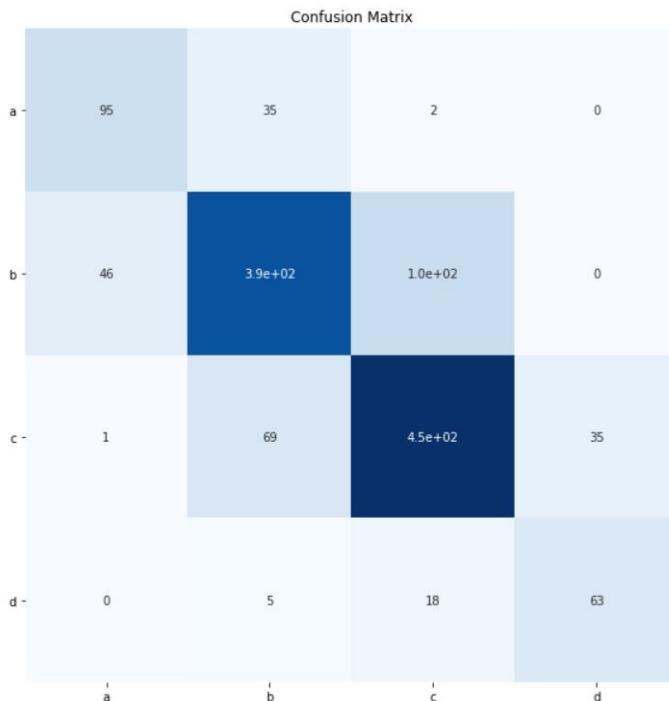


Figura 4.4.2: Matriz de confusión para la clasificación con SVC.

Capítulo 5

Conclusiones

Enfocándonos en los objetivos planteados podemos enumerar las siguientes conclusiones:

- Se pudo comprobar que existe una relación entre la textura de la imagen y la densidad radiológica según la clasificación dada por BI-RADS.
- Se alcanzaron muy buenos resultados en el algoritmo de segmentación automática de la imagen.
- Se logró obtener un modelo que permite una clasificación de la densidad radiológica de la mama con un promedio de aciertos de 0.76 para cada clase. Esto da una idea de la cantidad de coincidencias entre el sistema desarrollado y la clasificación subjetiva del médico especialista.
- Si nos remitimos a la bibliografía de referencia, se puede verificar que los resultados obtenidos se encuentran dentro del mismo orden que el alcanzado por nuestro modelo. Sin embargo, la comparación no puede ser directa debido a que las bases de datos utilizadas en dichos trabajos han sido preparadas especialmente para investigación, esto significa que han sido revisadas repetidamente por diferentes especialistas para disminuir el error asociado a la subjetividad humana. Otro punto a destacar del presente trabajo es que se han incluido tanto los casos patológicos como los normales y las mamas con implantes, que no han sido contempladas en otros trabajos..
- Se pudo comprobar que, al menos con las imágenes con las que contamos, el hecho de aumentar la cantidad de características, aumentando también los tiempos de cómputo, no resultó en un incremento de la cantidad de aciertos ya que al analizar separadamente conjuntos de características se lograron obtener resultados similares.

Capítulo 6

Trabajos futuros

1. Evaluar el desempeño en la clasificación balanceando la cantidad de ejemplos por clases.
2. Evaluar el desempeño en la clasificación agregando una red neuronal convolucional pre entrena-
da para extraer un mayor número de características.
3. Indagar sobre nuevas características que sean capaces de entregar información para la clasifi-
cación.
4. Disminuir tiempos de cómputo.
5. Desarrollar un sistema automático de bajo costo para servicios de diagnóstico por imágenes de
la región.

Bibliografía

- [1] Michel Latarjet and Alfredo Ruiz Liard. *Anatomía humana*, volume 2. Ed. Médica Panamericana, 2004. 3
- [2] Frank H Netter. *Atlas of Human Anatomy E-Book*. Elsevier Health Sciences, 2017. VIII, 4
- [3] Ministerio de Salud de la Nación. [http://www.msal.gov.ar/inc/acerca-del-cancer/cancer-de-mama/.](http://www.msal.gov.ar/inc/acerca-del-cancer/cancer-de-mama/>.) Última vez visitada Octubre de 2018. 5
- [4] Organización Mundial de la Salud. <http://www.who.int/>, Última vez visitada Octubre de 2018. 5
- [5] D'Orsi CJ Bassett LW Sickles, EA. *BI-RADS del ACR: Mamografía*. Reston, VA, Colegio Estadounidense de Radiología, 2013. 5, 22, 26
- [6] Stewart C. Bushong. Harcourt. *Manual de radiología para técnicos. Física, biología y protección radiológica*. 6 edition, 1999. VIII, VIII, VIII, VIII, VIII, 6, 8, 9, 10, 11, 14
- [7] Mamografía convencional. <http://kingsarai123.blogspot.com/2015/11/mamografia-convencional.html>, Última vez visitada Octubre de 2018. VIII, 7
- [8] Judith Kelly Peter Hogg and Claire Mercer. *Digital Mammography. A Holistic Approach*. Switzerland, 2015. VIII, 12, 18, 21, 22
- [9] R. Torres M. Chevalier. Mamografía digital. *Revista Física Médica*, 2010. VIII, VIII, 15, 16, 17
- [10] Wolfe J. J Natl Cancer Inst Byrne C, Schairer C. Mammographic features and breast cancer risk: effects with time, age, and menopause status. 1995. 20
- [11] Wolfe JN. AJR Am J Roentgenol. Breast patterns as an index of risk for developing breast cancer. 1976. 20, 22
- [12] Vogt K. Lancet Oncol Boyd NF, Rommens JM. Mammographic breast density as an intermediate phenotype for breast cancer. 2005. 20
- [13] Martin L. Sun L. Stone J. Fishell E. Boyd N.F., Guo H. Mammographic density and the risk and detection of breast cancer. *New England Journal of Medicine*. 22

- [14] Innolitics. <https://dicom.innolitics.com/ciods/digital-mammography-x-ray-image>, 2018. 26, 91
- [15] Enrique Alegre Gutiérrez. Procesamiento digital de imágenes: Fundamentos y prácticas con matlab. *Universidad de León, Secretariado de Publicaciones y Medios Audiovisuales*,. VIII, IX, 27, 28, 29, 30, 31, 32, 36, 37, 38, 39
- [16] Visión Artificial. https://es.wikipedia.org/wiki/Visión_artificial Octubre de 2018. 29
- [17] Cómo filtrar el ruido de una máscara con OpenCV. <https://robologs.net/2015/07/26/como-filtrar-el-ruido-de-una-mascara-con-opencv/>, Última vez visitada Octubre de 2018. VIII, 33
- [18] Teledetección. <http://www.aet.org.es/revistas/revista17/aet17-04.pdf>, Última vez visitada Octubre de 2018. VIII, 34
- [19] Técnicas de Filtrado. <https://www.um.es/geograf/sigmur/teledet/tema06.pdf>, Última vez visitada Octubre de 2018. IX, 35
- [20] Método Ridler Calvard y Otsu. <https://prezi.com/povqek3wduzr/metodo-ridler-calvar-y-metodo-otsu/>, Última vez visitada Octubre de 2018. 36
- [21] Método de Otsu. www.ugr.es/iloes/proyectos/matematicas/otsu.pptx, Última vez visitada Octubre de 2018. 37
- [22] Carmen Ximénez Gómez, Manuel Suero Suñe, and Juan Botella Ausina. *Análisis de datos en psicología I*. Ediciones Pirámide, 2014. 39
- [23] Diego Martín Mateos. Medidas de complejidad y de información como herramientas para el análisis de series temporales: aplicaciones al estudio de señales de origen electrofisiológicos. 2016. 41
- [24] Alceu Ferraz Costa, Gabriel Humpire-Mamani, and Agma Juci Machado Traina. An efficient algorithm for fractal analysis of textures. In *Graphics, Patterns and Images (SIBGRAPI), 2012 25th SIBGRAPI Conference on*, pages 39–46. IEEE, 2012. 42, 89
- [25] Robert M Haralick, K Shanmugam, Its'Hak Dinstein, et al. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, 3(6):610–621, 1973. IX, 43, 44, 91
- [26] Enrique J Carmona Suárez. Tutorial sobre máquinas de vectores soporte (svm). *Tutorial sobre Máquinas de Vectores Soporte (SVM)*, 2014. 46
- [27] Iñaki Inza y Pedro Larrañaga Abdelmalik Moujahid. Clasificadores k-nn. *Departamento de Ciencias de la Computación e Inteligencia Artificial Universidad del País Vasco–Euskal Herriko Unibertsitatea*, 2010. 53

- [28] BIOESTADISTICA (55 10536). Introducción a la regresión logística. *Departamento de Estadística Universidad Carlos III de Madrid*, 2010. 55
- [29] Random Forest Simple Explanation. <https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d>, Última vez visitada Octubre de 2018. IX, 57
- [30] Componentes principales. <https://www.mhe.es/universidad/cienciasmatematicas/pena/home/capitulo.pdf>, Última vez visitada Octubre de 2018. 57
- [31] Gustavo J. Meschino Adriana Antonelli and Virginia L. Ballarin. Cuantificación de la densidad mamaria mediante entropía de permutación en mamografías. *Laboratorio de Bioingeniería, Laboratorio de Procesamiento de Imágenes, Instituto de Investigaciones Científicas y Tecnológicas en Electrónica, ICYTE, UNMDP-CONICET, Argentina*, 2017. V, IX, X, 61, 86
- [32] Styliani Petroudi and Michael Brady. Breast density characterization using texton distributions. *33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA*, 2011. v, x, 62
- [33] E Fishell R A Jongk J W Byng, N F Boyd and M J Yaffe. Automated analysis of mammographic densities. 2013. v, x, 62
- [34] Juan Antonio Solves Llorens. Análisis de la densidad de mama asistido por ordenador. *Trabajo de investigación del Master de Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia*. v, x, 62, 63
- [35] N Karssemeijer. Automated classification of parenchymal patterns in mammograms. *University Hospital Nijmegen, Department of Radiology, PO Box 9101, 6500 HB Nijmegen, The Netherlands*, 7(14):355–377, 1998. v, x, 63, 82
- [36] X. Lladó C. Mata, J. Freixenet and A. Oliver. Texture descriptors applied to digital mammography. 2008. v, x, 63, 64

Apéndice A

DDSM (The Digital Database for Screening Mammography)

La base de datos digital para mamografías de detección (DDSM, por sus siglas en inglés) es un recurso para el uso de la comunidad de investigación en análisis de imágenes mamográficas. El apoyo principal para este proyecto fue una subvención del Programa de Investigación del Cáncer de Mama del Comando de Investigación Médica y Material del Ejército de los EE. UU. El proyecto DDSM es un esfuerzo de colaboración que involucra co-pis en el Hospital General de Massachusetts (D. Kopans, R. Moore), la Universidad del Sur de la Florida (K. Bowyer) y los Laboratorios Nacionales Sandia (P. Kegelmeyer).^{1 2}

DDSM está constituida por 2,620 estudios de mamografía de película escaneada (digitalizadas). Contiene casos normales, benignos y malignos con información de patología verificada. La escala de la base de datos, junto con la validación de la verdad fundamental, hace que el sea una herramienta útil en el desarrollo y prueba de los sistemas de soporte de decisiones. Las imágenes se proporcionan en formato DICOM (163 GB).

Finalmente, las anotaciones de ROI para las anomalías en el DDSM se proporcionaron para indicar una posición general de las lesiones, pero no una segmentación precisa para ellas. Por lo tanto, muchos investigadores deben implementar algoritmos de segmentación para la extracción precisa de características. Esto provoca una incapacidad para comparar directamente el rendimiento de los métodos o para replicar resultados anteriores.

Se proporciona además una tabla utilizada para implementar algoritmos de clasificación de patologías mamarias, con el fin de poder predecir la probabilidad de malignidad o benignidad del tumor. Esta tabla está constituida por:

- Densidad radiológica (1-4).

¹<http://marathon.csee.usf.edu/Mammography/Database.html>

²<https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM>

- Si posee patología, se indica si es una masa o una microcalcificación (cantidades).
- Forma de la región sospechosa.
- Clasificación BIRADS (0-6).