ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA
GRADO EN INGENIERÍA DE LA SALUD

PROCESAMIENTO Y ANÁLISIS DE MAMOGRAFÍAS PARA LA DETECCIÓN DE TUMORES

PROCESSING AND ANALYSIS OF MAMMOGRAMS FOR DETECTING TUMORS

Realizado por
DIEGO IGNACIO OSORIO FORTEA
Tutorizado por
ENRIQUE DOMÍNGUEZ MERINO
Departamento
LENGUAJES Y CIENCIAS DE LA COMPUTACIÓN

UNIVERSIDAD DE MÁLAGA
MÁLAGA, DICIEMBRE 2015

Fecha defensa:
El Secretario del Tribunal

Resumen

Este trabajo nace de una idea conjunta del desarrollo de un software capaz de trabajar con una máquina de rayos X para la toma de mamografías, y ser una herramienta para los radiólogos a la hora de analizar las imágenes y localizar zonas que pudieran ser sensibles de tener algún tipo de tumoración. No se trata de un software que pretenda suplantar la función de un radiólogo; es importante tener claro que el experto en radiología es el único capaz de analizar, comparar y tomar decisiones reales, y este software pretende simplificar esa tarea al máximo.

Este programa utiliza como recursos para su funcionamiento una base de datos de mamografías procesadas anteriormente, a partir de las cuales entrena a la red neuronal para que sea capaz de clasificar mamografías nuevas y mostrar las posibles zonas críticas. Esta base de datos debe indicar a la red qué mamografías presentan algún tumor y dónde, para que la red aprenda a diferenciar entre zonas críticas y zonas no críticas.

Esta idea deriva en la realización de dos trabajos que se especializan en los dos ámbitos del proyecto, en primer lugar el procesamiento y el análisis de esas mamografías obtenidas de la máquina de rayos X, que da lugar al presente trabajo; y en segundo lugar, deriva en un trabajo paralelo que explica la obtención, el análisis y la clasificación de los datos que se desprenden de este proyecto utilizando redes neuronales, titulado “Detección y clasificación de tumores en mamografías a través de redes neuronales”; realizado por Rodrigo Culotta López.

Palabras claves

Mamografía, tumor, cáncer, procesamiento, detección, imagen, descriptores, software, Matlab

Abstract

This project stems from a joint idea to develop a software capable of working with an X-ray machine in order to capture mammograms, with the purpose of becoming a tool for radiologists when analysing the images and locating zones that could be prone to having a tumour. This program does by no means intend to supplant the function of a radiologist as it is important to bear in mind that the said doctors are the only ones who are able to analyse, compare and make real decisions, being the aim of this program to simply facilitate their job as much as it possibly can.

A database of mammograms previously processed is used by this program as a resource for its performance. It trains the neural network due to its ability to classify new mammograms and show the possible critical zones. This database must indicate which mammograms have a tumour as well as its location, so that the net learns to differentiate between critical and non-critical zones.

This idea derives in the elaboration of two documents that specialise in both aspects of the project: one dealing with the processing and the analysis of the mammograms obtained from the X-ray machine, originating the present document; and the other leading to a parallel document which explains the obtention, analysis and classification of the information which comes from this part of the project that utilises neural networks, named “Detection and classification of tumours in mammograms using neural networks”; made by Rodrigo Culotta López.

Keywords

Mammogram, tumor, cancer, processing, detection, image, descriptors, software, Matlab

Índice de contenidos

1. Introducción

1.1. Motivación

Hoy en día, la detección de posibles tumores a través de la inspección de imágenes de rayos X por parte de un médico se ha convertido en un proceso rutinario. El médico, utilizando su conocimiento y capacidad de observación, puede clasificar las imágenes biomédicas atendiendo al criterio de si presentan o no tumores.

Este trabajo de detección y clasificación es realizado por un humano. Esto puede dar lugar a que se cometan ciertos errores durante uno de los dos procesos. En el caso de que estos médicos recibieran una ayuda complementaria, como podría ser la ayuda de un software para realizar esa labor, se reducirían los posibles errores que éstos podrían cometer ante dicha tarea.

De aquí surge nuestra idea, centrada en un software capaz de realizar esas tareas de la forma más eficiente posible y que pueda ser utilizado por el médico de forma complementaria para disminuir el error humano.

El estudio derivará en dos trabajos fin de grado, estando el presente trabajo centrado en las fases de localización, detección y estudio de características de las imágenes; y uno paralelo a éste, centrado en la parte de inteligencia artificial, utilizando redes neuronales para el procesamiento y clasificación de las imágenes según posean zonas con tumor o no.

Algunos problemas como el diagnóstico de enfermedades y el reconocimiento de imágenes se han intentado resolver usando técnicas de clasificación. El cáncer de mama es un cáncer muy común y grave en la mujer, siendo la mamografía uno de los métodos más utilizados para detectarlo [1].

Técnicas estadísticas y técnicas de inteligencia artificial se han utilizado para predecir el cáncer de mama en numerosas investigaciones. El objetivo de estas técnicas de identificación es la clasificación de pacientes en ‘benigno’ si no tienen cáncer de mama o ‘maligno’ si hay pruebas de que tienen cáncer de mama. Por tanto, el diagnóstico de cáncer de mama es un problema muy discutido entre los problemas de clasificación.

Existen muchas técnicas para predecir y clasificar este tipo de cáncer:

- Aragonés, Ruiz, Jiménez, Pérez y Conejo utilizaron una red neuronal combinada y varios modelos de árboles de decisión para el diagnóstico [2].
- Choua et al utilizó redes neuronales artificiales y regresión multivariable adaptativa para la clasificación del cáncer de mama [3].
- Ryua et al utilizó técnicas de separación isotópica para la predicción del cáncer de mama [4].
- Sahan, Polat, Kodaz y Günes crearon un nuevo método híbrido basado en un sistema artificial inmune y un algoritmo k-nn para dicho diagnóstico [5].
- Úbeyli clasificó los datos de cáncer de mama en Wisconsin usando redes neuronales multicapa, redes neuronales combinadas, redes neuronales probabilísticas, redes neuronales recurrentes y máquinas vectoriales [6].

En la actualidad, son muchas las empresas que comercializan con sistemas de mamografías y, sin embargo, son pocas las principales empresas del sector que cuentan con equipos que posean un software integrado para la detección de posibles zonas tumorales; como es el caso de Fujifilm [7]. Sin embargo, otras empresas de referencia en el sector, como General Electric, Siemens o Philips, aún no cuentan con herramientas de este tipo [8] [9] [10].

1.2. Objetivos

Queremos diseñar un software que sea capaz de realizar un procesado de las mamografías para así poder detectar la zona del tumor utilizando para ello la ayuda de redes neuronales entrenadas a partir de una base de datos de mamografías para dicho fin.

El procesado consistirá en el cambio de contraste de las imágenes para aumentar la visibilidad de la zona de interés de estudio. Estas zonas se detectarán y se analizarán con la ayuda de descriptores morfológicos de distinto tipo que describirán ciertas características importantes de dicha zona.

Para ello, tendremos que utilizar descriptores de diferente tipo: de área, de forma, momentos, entre otros; de manera que se llevaría un proceso reiterativo en el que tendremos que ir probando descriptores diferentes para ver cuál es el que mejor encaja con el objetivo propuesto en el proyecto.

Obteniendo estas características, podremos utilizarlas para entrenar a la red neuronal, que será la encargada de la detección de la zona tumoral a través de los datos proporcionados por estos descriptores.

Por tanto, nuestro objetivo final es desarrollar un software capaz de detectar todos los cánceres y clasificarlos correctamente; más concretamente, queremos obtener 0 cánceres clasificados como zonas normales aunque para ello aumente el número de zonas no cancerígenas clasificadas como cánceres, ya que es el radiólogo quien tiene la última palabra y analizará las zonas destacadas, las cuales no quieren decir que eso sea un cáncer, sino que presenta ciertas condiciones para poder serlo; pero lo esencial es que no se pase por alto ningún cáncer.

1.3. Estructura de la memoria

- Recolección de mamografías.
- Eliminación de mamografías no válidas.
- Localización de tumores malignos conocidos y aislamiento de los mismos.
- Procesado de la imagen
- Detección de zonas potencialmente tumorales y aislamiento de las mismas.
- Estudio de los descriptores morfológicos de cada fragmento.
- Resultados

1.4. Tecnologías a utilizar

Este trabajo basa únicamente su realización en el software matemático Matlab, más concretamente en sus herramientas de procesado de imágenes y cálculos con matrices.

2. Recolección de mamografías

El primer paso consiste en obtener muestras de mamografías para comenzar con el desarrollo del software. Necesitaremos imágenes biomédicas de mamas que presenten cáncer y mamas que no lo presenten. Lo más importante de este procedimiento es obtener una cantidad proporcionada de ambos tipos de imágenes para poder realizar el entrenamiento de la red con el mayor número de diferentes muestras posibles.

Para ello, hemos utilizado la base de datos de mamografías digitales de la *South Florida University* [11]. En dicha base de datos podemos encontrar cientos de casos de diferentes mujeres con y sin cáncer, presentándose las mamografías de cada caso, la localización del tumor en el caso de que lo presenten y, si presentan cáncer, un archivo anexo con formato .OVERLAY, con datos acerca de las mamografías de ese caso tales como el pixel central del tumor, si el tumor que presenta es maligno o no, la forma del tumor, etc.

La página web nos presenta diferentes conjuntos de casos clasificados dependiendo de varios aspectos (Figura 01):

- Imágenes con cáncer o normales.
- El tipo de escáner utilizado.
- La resolución de la imagen.
- El número de BITS de la imagen.

VOLUME	CASES	SIZE	SCANNER	BITS	RESOLUTION	THUMBNAILS	NOTES	AVAILABILITY
normal_01	111	5.8 GB	DBA	16	42 microns	thumbnails	notes	ftp
normal_02	117	6.6 GB	DBA	16	42 microns	thumbnails	notes	ftp
normal_03	38	4.1 GB	DBA	16	42 microns	thumbnails	notes	ftp
normal_04	57	5.1 GB	DBA	16	42 microns	thumbnails	notes	ftp
normal_05	47	4.3 GB	DBA	16	42 microns	thumbnails	notes	ftp
normal_06	60	5.5 GB	DBA	16	42 microns	thumbnails	notes	ftp
normal_07	78	6.2 GB	HOWTEK	12	43.5 microns	thumbnails	notes	ftp
normal_08	27	2.8 GB	HOWTEK	12	43.5 microns	thumbnails	notes	ftp
normal_09	59	4.9 GB	LUMYSIS	12	50 microns	thumbnails	notes	ftp
normal_10	23	2.1 GB	LUMYSIS	12	50 microns	thumbnails	notes	ftp
normal_11	58	6.1 GB	HOWTEK	12	43.5 microns	thumbnails	notes	ftp
normal_12	20	2.2 GB	HOWTEK	12	43.5 microns	thumbnails	notes	ftp
cancer_01	69	3.9 GB	LUMISYS	12	50 microns	thumbnails	notes	ftp
cancer_02	88	5.7 GB	LUMISYS	12	50 microns	thumbnails	notes	ftp
cancer_03	66	6.0 GB	DBA	16	42 microns	thumbnails	notes	ftp
cancer_04	31	2.8 GB	DBA	16	42 microns	thumbnails	notes	ftp
cancer_05	83	6.6 GB	LUMISYS	12	50 microns	thumbnails	notes	ftp
cancer_06	56	6.3 GB	HOWTEK	12	43.5 microns	thumbnails	notes	ftp
cancer_07	52	6.1 GB	HOWTEK	12	43.5 microns	thumbnails	notes	ftp
cancer_08	55	6.0 GB	HOWTEK	12	43.5 microns	thumbnails	notes	ftp
cancer_09	81	6.5 GB	LUMISYS	12	50 microns	thumbnails	notes	ftp
cancer_10	59	6.6 GB	HOWTEK	12	43.5 microns	thumbnails	notes	ftp
cancer_11	59	5.9 GB	HOWTEK	12	43.5 microns	thumbnails	notes	ftp
cancer_12	80	6.8 GB	HOWTEK	12	43.5 microns	thumbnails	notes	ftp

Figura 01. Formato de presentación de los conjuntos de mamografías en la Web.

Dentro de cada conjunto de datos, se nos muestran los diferentes casos en una interfaz sencilla. Dentro de caso, podemos visualizar cada mamografía que lo compone con sus datos de interés remarcados en diferente color, ya sea para mostrar los tumores que presenta la imagen o diferentes zonas de interés (Figura 02).

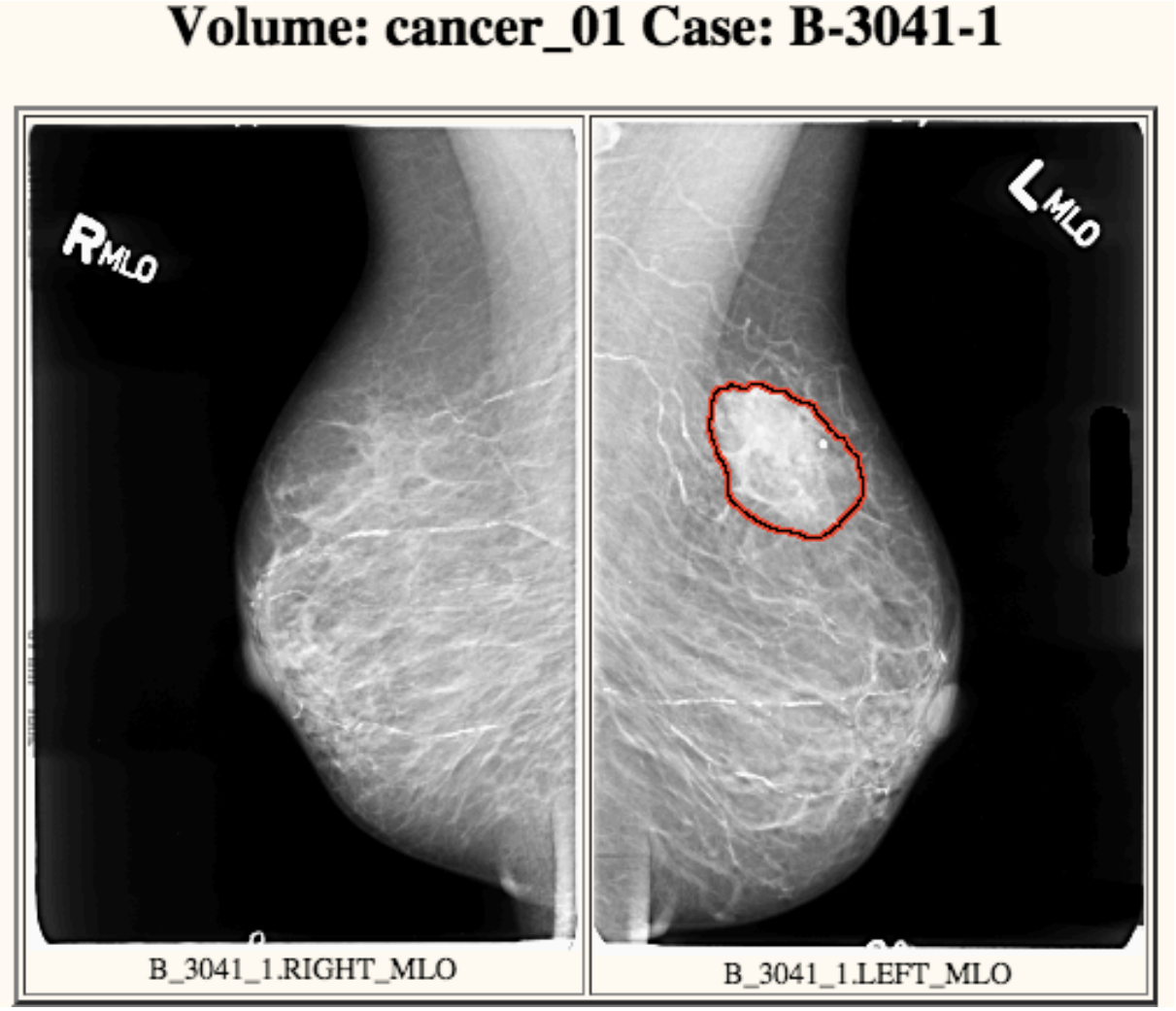
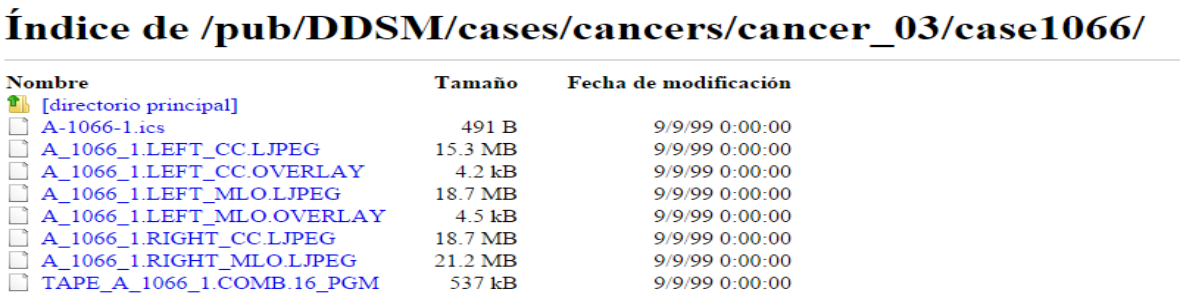


Figura 02. Visualización previa de las mamografías "B_3041_1.RIGHT.MLO" Y "B_3041_1.LEFT_MLO", con su zona tumoral remarcada en el caso de la segunda.

Todas las imágenes de mamografías se presentan en formato .LJPEG, formato que MATLAB puede reconocer y abrir, no sin cometer algunos errores de procesamiento. Existen varios métodos para convertir este formato, propio de esta universidad, en un formato más estándar de imagen, pero todos estos métodos se encuentran actualmente obsoletos.

La última columna de la tabla de imágenes es AVAILABILITY, desde donde se pueden descargar las mamografías con sus respectivos archivos OVERLAY si los tienen (Figura 03).



Índice de /pub/DDSM/cases/cancers/cancer_03/case1066/

Nombre	Tamaño	Fecha de modificación
[directorio principal]		
A-1066-1.ics	491 B	9/9/99 0:00:00
A_1066_1.LEFT_CC.LJPEG	15.3 MB	9/9/99 0:00:00
A_1066_1.LEFT_CC.OVERLAY	4.2 kB	9/9/99 0:00:00
A_1066_1.LEFT_MLO.LJPEG	18.7 MB	9/9/99 0:00:00
A_1066_1.LEFT_MLO.OVERLAY	4.5 kB	9/9/99 0:00:00
A_1066_1.RIGHT_CC.LJPEG	18.7 MB	9/9/99 0:00:00
A_1066_1.RIGHT_MLO.LJPEG	21.2 MB	9/9/99 0:00:00
TAPE_A_1066_1.COMB.16_PGM	537 kB	9/9/99 0:00:00

Figura 03. Apartado AVAILABILITY.

3. Eliminación de mamografías no válidas

Una vez realizada la primera tarea de recolección de imágenes, nuestro siguiente objetivo es conseguir dos conjuntos de imágenes (con cáncer y normales) formados por imágenes que cumplan unas condiciones mínimas que aseguren el correcto procesamiento de las mismas.

El principal problema que nos encontramos es el formato “ljpeg”, ya que Matlab puede procesarlo pero cometiendo errores que pueden derivar en dos tipos de imágenes:

- Imágenes ruidosas.
- Imágenes que presentan un triángulo en la zona superior derecha.

Este problema con el formato “ljpeg” se ha sido intentando subsanar de diferentes maneras: desde herramientas que la propia universidad proporcionaba, hasta softwares desarrollados por otros investigadores que han trabajado con esta base de datos de imágenes; pero los softwares que hemos probado para convertir este formato “ljpeg” en el formato estándar “jpeg” no han dado resultado debido a diversos fallos relacionados con la antigüedad de estos programas o métodos, como es el caso de un software que desarrolló el Dr. Chris Rose, de la Universidad de Manchester [12].

Tras estos fallidos intentos de conversión de formato, decidimos seguir adelante con la lectura que realizaba Matlab de “ljpeg” y trabajar con aquellas imágenes que no se veían influenciadas por el error introducido al ser procesadas, es decir, aquellas imágenes donde la zona de interés estaba claramente diferenciada de la zona errónea.

En el caso de las imágenes ruidosas, quedarán automáticamente descartadas ya que son inservibles (Figura 04); sin embargo, en el caso de las imágenes con un triángulo, tenemos que analizar cada una de ellas y comprobar si nuestra zona de interés y estudio, la zona cancerígena, está completamente presente en la imagen (Figura 05); o si de lo contrario solo aparece parcialmente (Figura 06), supondría su descarte para el conjunto de imágenes de cáncer, pues nuestro objetivo es aislar las zonas tumorales puras para su posterior análisis, y una zona parcial derivaría en datos incompletos y confusos.

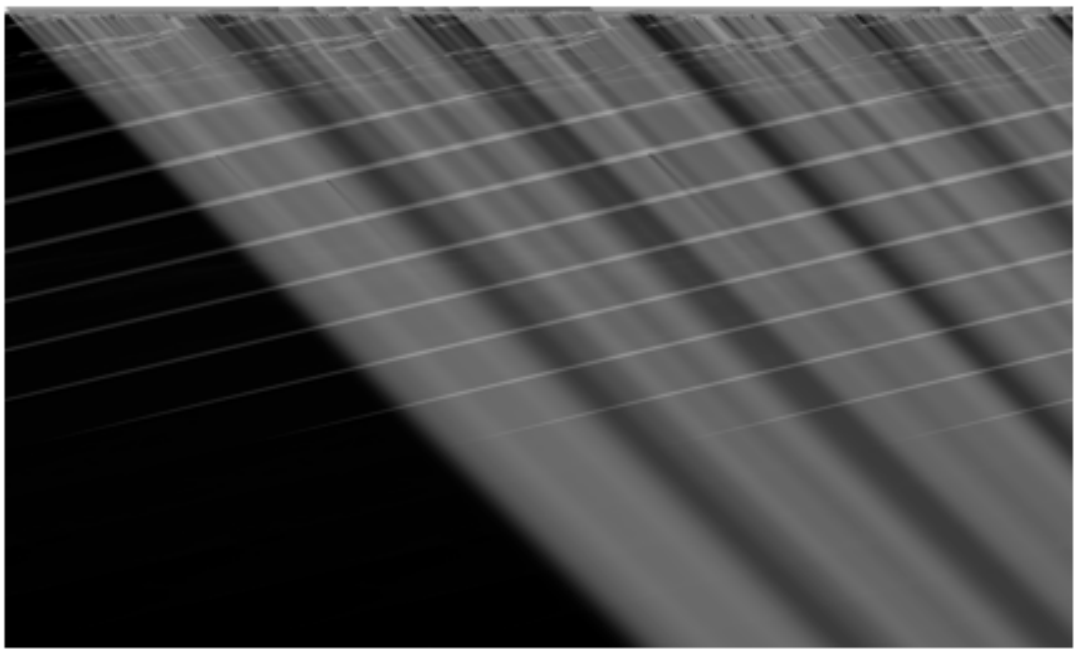


Figura 04. Mamografía “A_1014_1.LEFT_CC”. Ejemplo de imagen ruidosa debido a una lectura incorrecta.

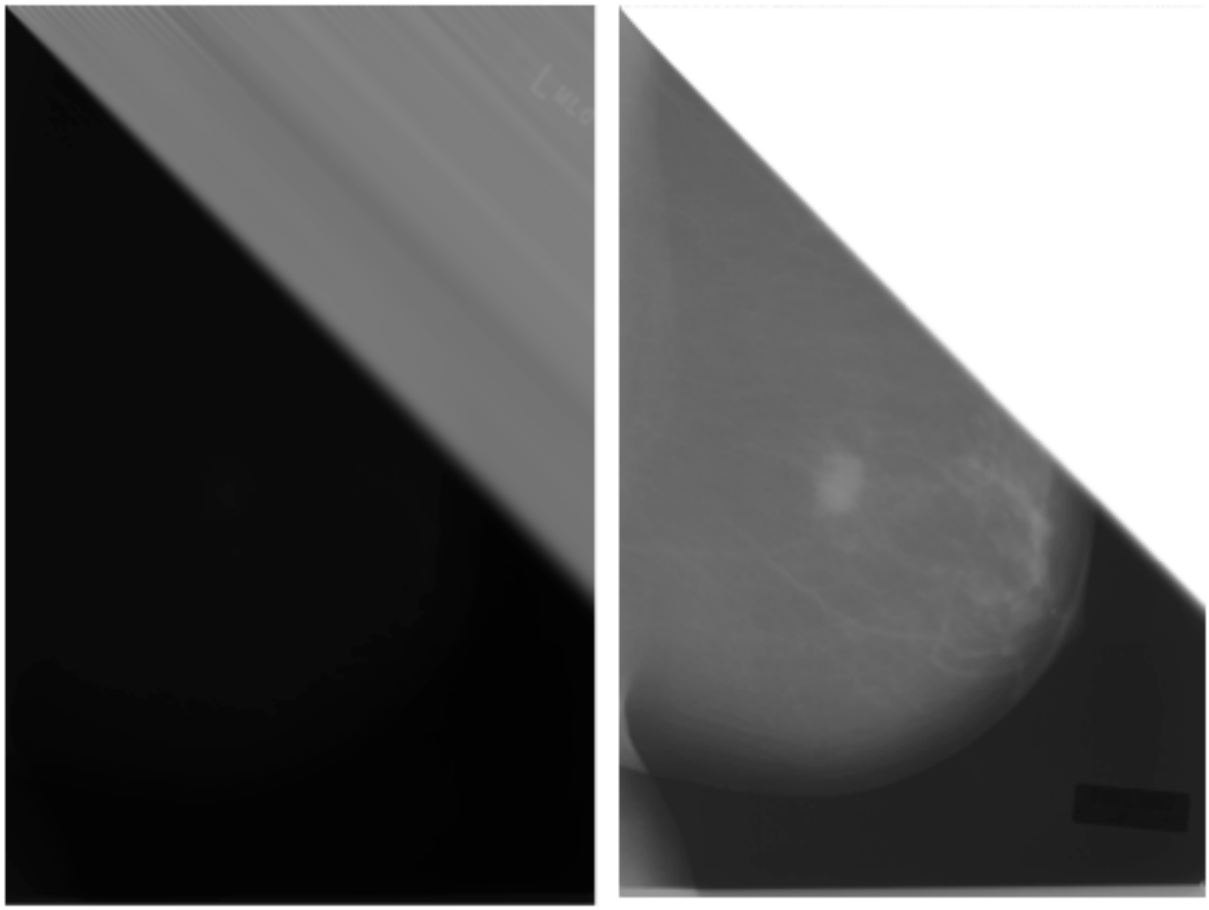


Figura 05. Mamografía “C_0037_1.LEFT_MLO”. Ejemplo de lectura correcta y cáncer visible, por lo que esta imagen entra dentro del conjunto de imágenes de cáncer con el que trabajaremos en adelante.

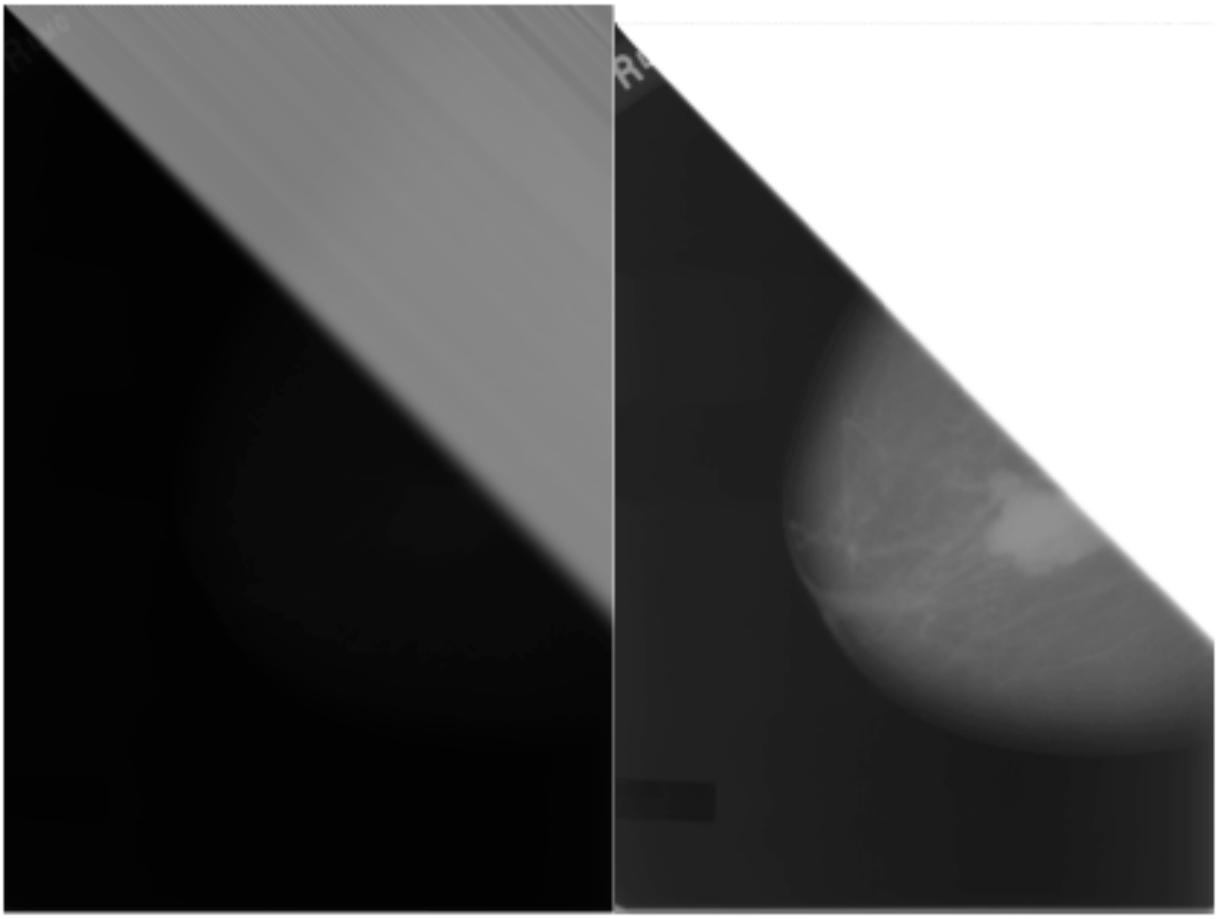


Figura 06. Mamografía “C_0011_1.RIGHT_MLO”. Ejemplo de lectura correcta pero inválida para nuestro estudio debido a que el cáncer se muestra parcialmente. A la izquierda se muestra la imagen original, y a la derecha la misma imagen modificado su contraste para una correcta visualización de la zona de interés.

De igual manera que ocurre con las imágenes con cáncer, las imágenes normales también presentan los dos tipos de errores (Figura 07). Sin embargo, las imágenes normales presentan un amplio abanico de ruido; condición que utilizaremos como filtro para la selección de dichas imágenes, en función del ruido presente y de la relevancia que suponemos que tendrá, seleccionaremos las imágenes que nos aseguren pocos o ningún inconveniente en el procesamiento.

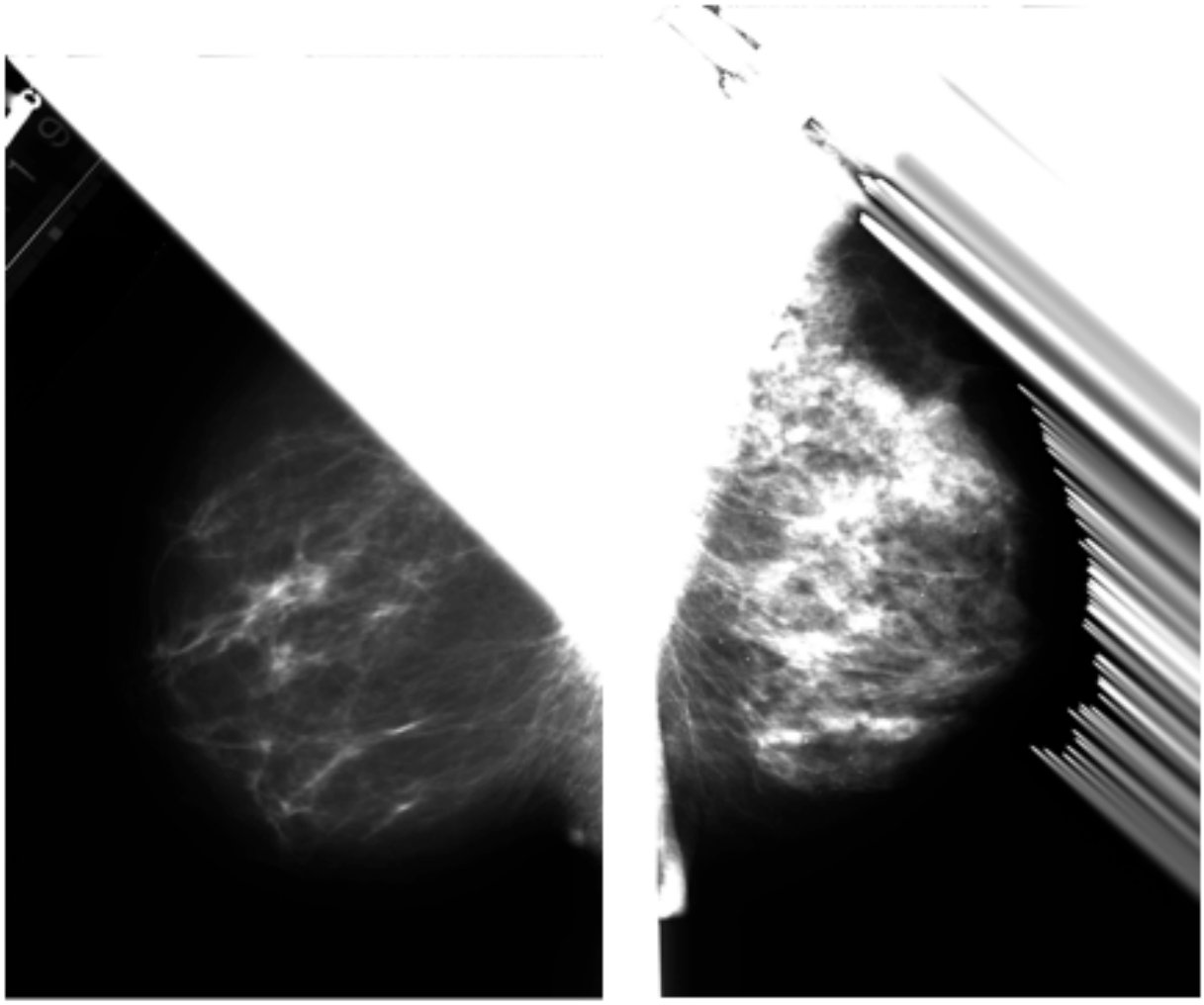


Figura 07. A la izquierda, la mamografía “A_0201_1.LEFT_MLO”; a la derecha la mamografía “C_1033_1.RIGHT_MLO”. La primera se muestra sin ruido provocado por la lectura del formato “ljpeg” y es una buena imagen para nuestro conjunto de imágenes normales. Por el contrario, la segunda imagen presenta mucho ruido y queda descartada como imagen para dicho conjunto.

4. Localización de tumores malignos conocidos y aislamiento de los mismos

Nuestra primera tarea es detectar y aislar los tumores de las mamografías que posean alguno. Para conocer esa información, contamos con la ayuda de los archivos OVERLAY; toda aquella imagen que vaya acompañada de este archivo significa que presenta algún tumor.

Los archivos OVERLAY contienen diferente tipo de información sobre la imagen (Figura 08), como puede ser el número de anormalidades que presenta, el tipo de patología, y lo que es más importante para nosotros, aparece el contorno de la zona tumoral en la última fila. Los dos primero valores son las coordenadas desde donde comienza el contorno, y los siguientes números indican la dirección en la que va creciendo siguiendo el siguiente patrón:

7	0	1
6	X	2
5	4	3

Sin embargo, para que el algoritmo de crecimiento de regiones [13] sea óptimo, hemos cambiado las coordenadas que indican el inicio del borde, y hemos colocado las coordenadas del centro de la mancha, ya que facilita el crecimiento en todas las direcciones; obteniendo unos mejores resultados.

```
TOTAL_ABNORMALITIES 1
ABNORMALITY 1
LESION_TYPE CALCIFICATION TYPE PUNCTATE DISTRIBUTION SEGMENTAL
ASSESSMENT 4
SUBTLETY 2
PATHOLOGY MALIGNANT
TOTAL_OUTLINES 4
BOUNDARY
1195 2243 7 7 7 7 7 7 7 7 0 0 0 0 0 0 0 0 7 7 7 7 7 7 7 7 0 0 0
```

Figura 08. Contenido del archivo OVERLAY. Los dos primeros números de la última fila indican las coordenadas de la semilla.

Para realizar esta detección, en primer lugar modificamos el contraste de la imagen para ampliar los niveles de grises comprendidos entre 0 y 0,1, utilizando la función “imadjust” [14], donde se encuentra la zona de interés.

Posteriormente, tenemos que acceder al archivo OVERLAY, si existe, lo cual indica que la imagen actual tiene algún tumor, y utilizamos la coordenada que indica el fichero para realizar un crecimiento de regiones. Este algoritmo se basa en crecer desde una semilla en función de los niveles de grises de los pixeles de su alrededor. Si la diferencia entre el valor del pixel actual y la del pixel colindante es menor que un umbral que previamente le hemos indicado a la función, la zona crece abarcando ahora el nuevo pixel. Este proceso continua hasta que no hay más píxeles colindantes que cumplan la condición previamente explicada.

La función nos devuelve una imagen del mismo tamaño que la original. Sin embargo, es una imagen en blanco y negro, en la que solo aparece en blanco la zona que ha surgido tras el crecimiento total desde la semilla (Figura 09).

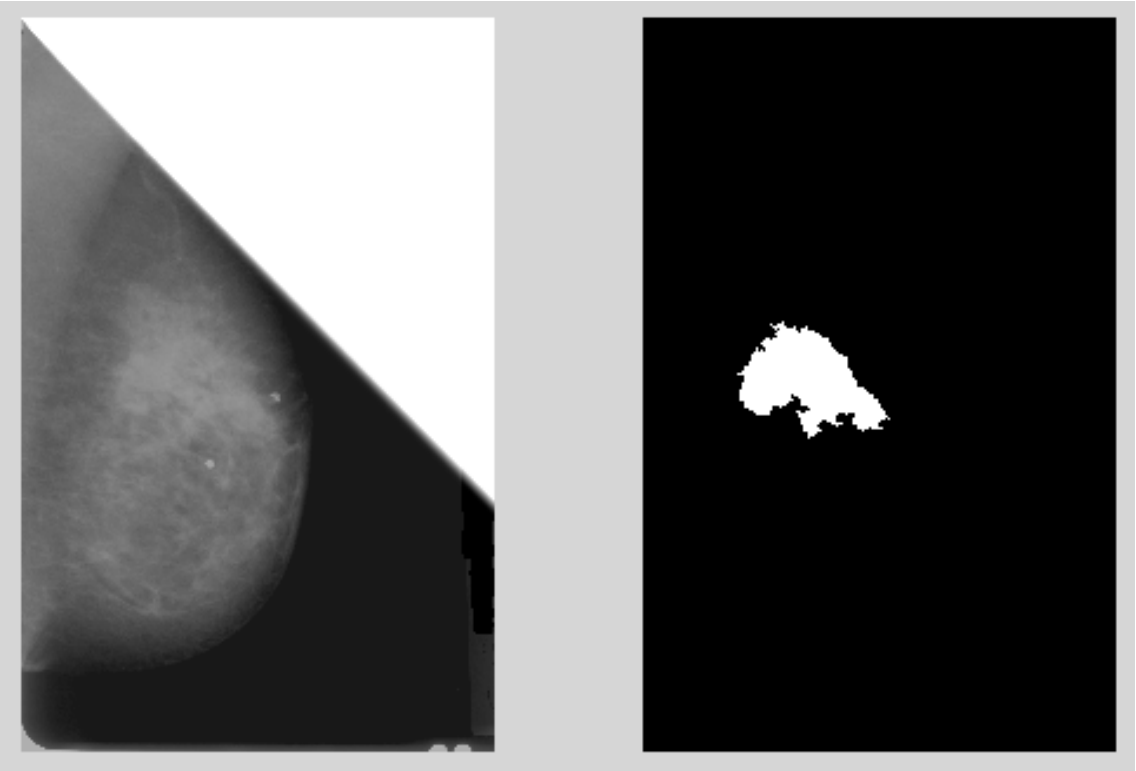


Figura 09. A la izquierda la imagen original con el contraste básico para su visualización clara. A la derecha la mancha obtenida por el algoritmo de crecimiento de regiones a partir de la semilla indicada en el archivo OVERLAY.

Una vez detectado el tumor, solo nos queda aislarlo de la imagen utilizando una función “recorte”. Esta función simplemente extrae el rectángulo mínimo que contiene a la mancha en su interior (Figura 10).

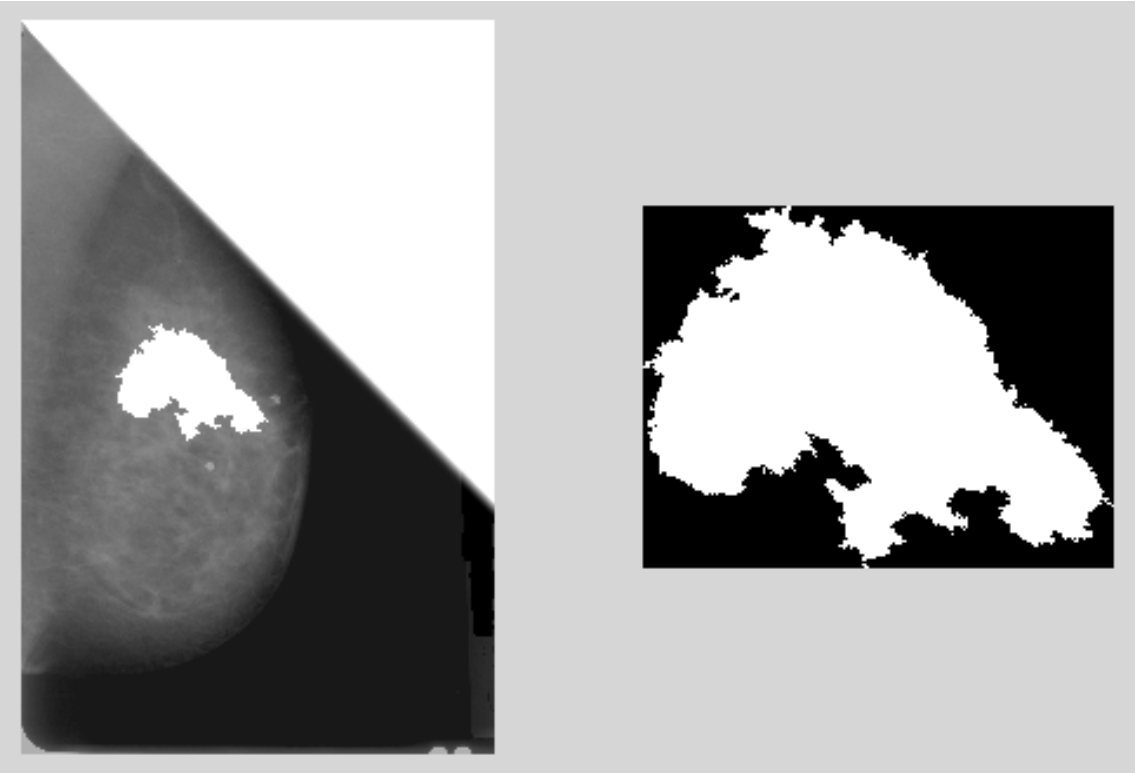


Figura 10. A la izquierda aparece la mancha obtenida sobre la imagen original, y a la izquierda aparece la mancha aislada con la función “recorte”.

Este proceso nos permitirá indicarle a la red que manchas son tumor, de esta manera sabrá que los datos que esta obteniendo de los descriptores morfológicos corresponden a un tumor. Este hecho le ayudará a aprender para que, ante nuevas manchas, en función de los valores de los descriptores, pueda clasificarlas en tumorales o no tumorales.

5. Procesado de la imagen

El procesamiento de las imágenes es parte más importante de este trabajo. Para poder obtener las manchas candidatas de todas las imágenes, es esencial que el procesamiento previo sea correcto y asegure que todas las zonas críticas estén destacadas y sean fáciles de detectar en el siguiente paso.

Para poder mostrar las etapas de este procesado, en adelante tomaremos dos imágenes como imágenes de estudio. Para establecer un criterio en la selección de estas dos imágenes, hemos seleccionado las imágenes cuyos valores medios de nivel de gris son el mayor y el menor, respectivamente (Figura 11). En adelante:

A = imagen con el valor de intensidad medio más alto
B = imagen con el valor de intensidad medio más bajo

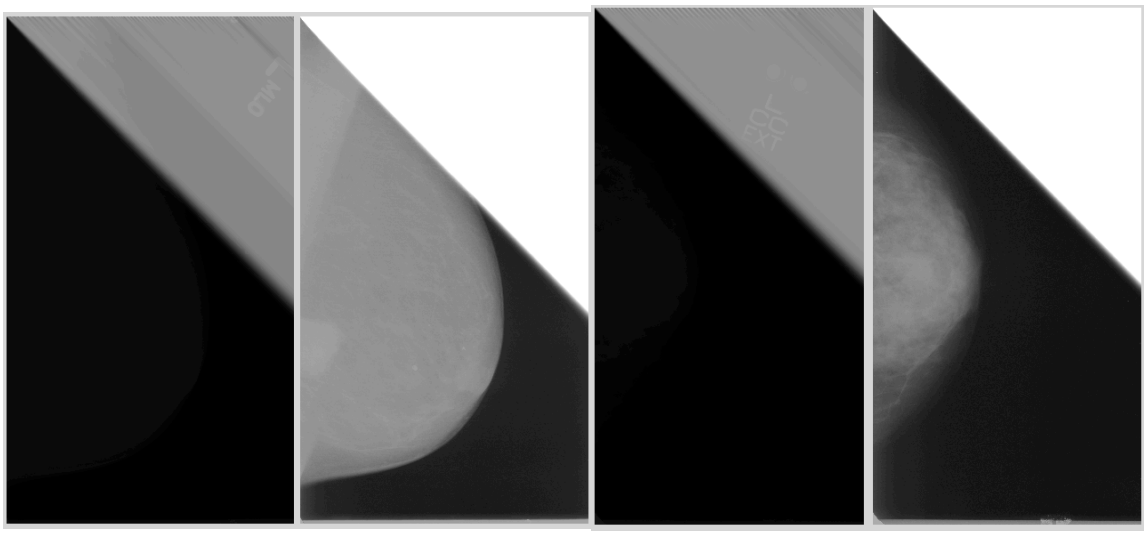


Figura 11. A la izquierda, nuestra imagen ejemplo A. A la derecha, nuestra imagen ejemplo B. La primera imagen de cada pareja corresponde a la imagen que se lee del fichero, y la segunda imagen es la misma pero ampliados los niveles de grises del intervalo [0;0.1] a [0;1].

5.1. Eliminación del triángulo

El primer paso del procesamiento es eliminar de todas las imágenes el triángulo introducido por la errónea lectura del formato “ljpeg”. Para ello, como hicimos en el apartado anterior para detectar los cánceres conocidos, utilizaremos el algoritmo de crecimiento de regiones.

Sabiendo que el triángulo siempre está situado en la misma zona, colocamos una semilla fija para todas las imágenes en la esquina superior derecha. El resultado del crecimiento de regiones es la imagen en negro con prácticamente la totalidad del triángulo en blanco. Restándole a la imagen de estudio la obtenida con la semilla, conseguimos realizar la eliminación (Figura 12).

Esta eliminación completa del polígono es imposible ya que existe una zona de transición de niveles de grises entre la zona de la imagen real y el triángulo. Más adelante, se abordará el proceso para eliminar estas líneas que puedan quedar.

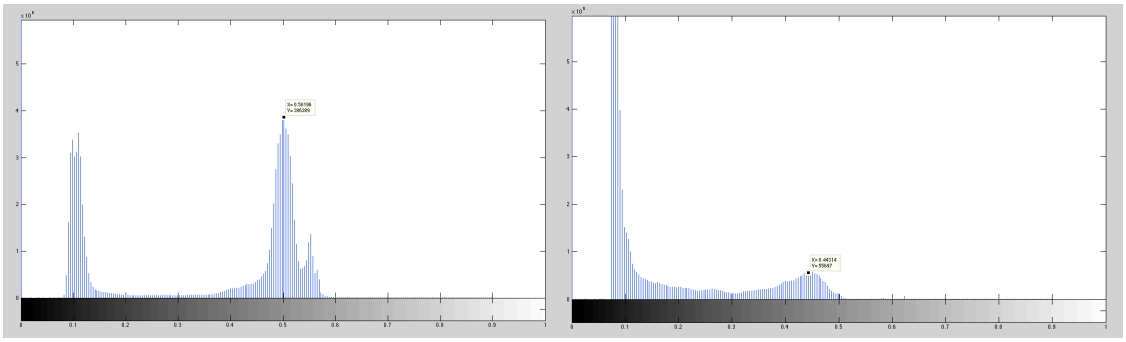


Figura 12. A la izquierda, el histograma de nuestra imagen ejemplo A (Figura 15, primera columna). A la derecha, el histograma de nuestra imagen ejemplo B (Figura 16, primera columna).

Como podemos comprobar en los histogramas anteriores, los valores de grises de las imágenes varían mucho, lo cual complica mucho nuestro estudio. Este hecho nos obliga a realizar otro tipo de procesado en las imágenes para homogeneizarlas.

5.2. Normalización

Este proceso es vital para poder tratar de igual manera a todas las imágenes y ser capaces de trabajar en unos rangos de niveles de grises determinados. La normalización se realiza gracias al comando “histeq” [15], lo que nos permite obtener imágenes más similares y con unos niveles de grises más homogéneos (Figura 13).

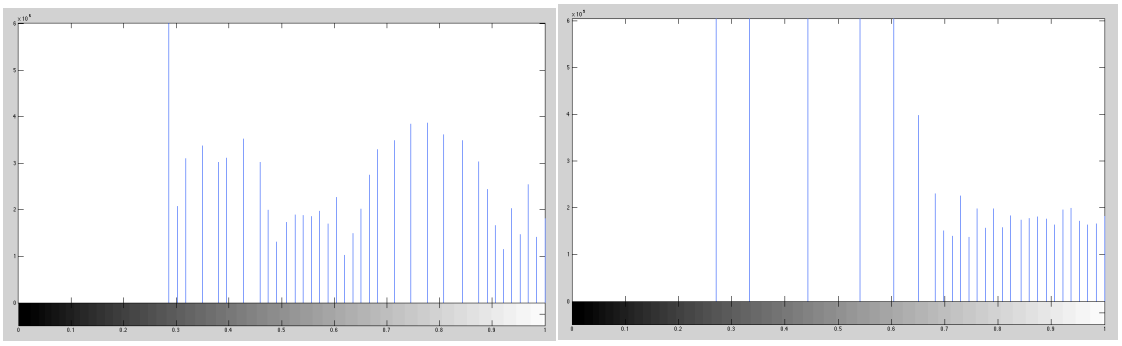


Figura 13. A la izquierda, el histograma de nuestra imagen ejemplo A (Figura 15, segunda columna). A la derecha, el histograma de nuestra imagen ejemplo B (Figura 16, segunda columna).

5.3. Ajuste del contraste

Una vez normalizadas las imágenes, nuestra tarea es seleccionar el rango de niveles de grises relevantes para nuestro estudio; los niveles de grises correspondientes a las zonas más claras de la imagen, utilizando un valor mínimo de corte lo suficientemente bajo como para tener la seguridad de que cualquier posible cáncer pase el filtro. Podemos comprobar que una vez realizado este proceso, los valores de grises de las imágenes son similares (Figura 14).

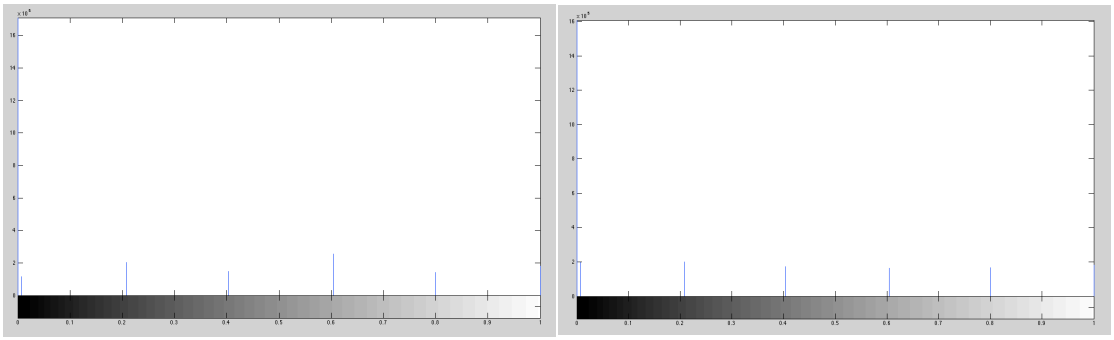


Figura 14. A la izquierda, el histograma de nuestra imagen ejemplo A (Figura 15, tercera columna). A la derecha, el histograma de nuestra imagen ejemplo B (Figura 16, tercera columna).

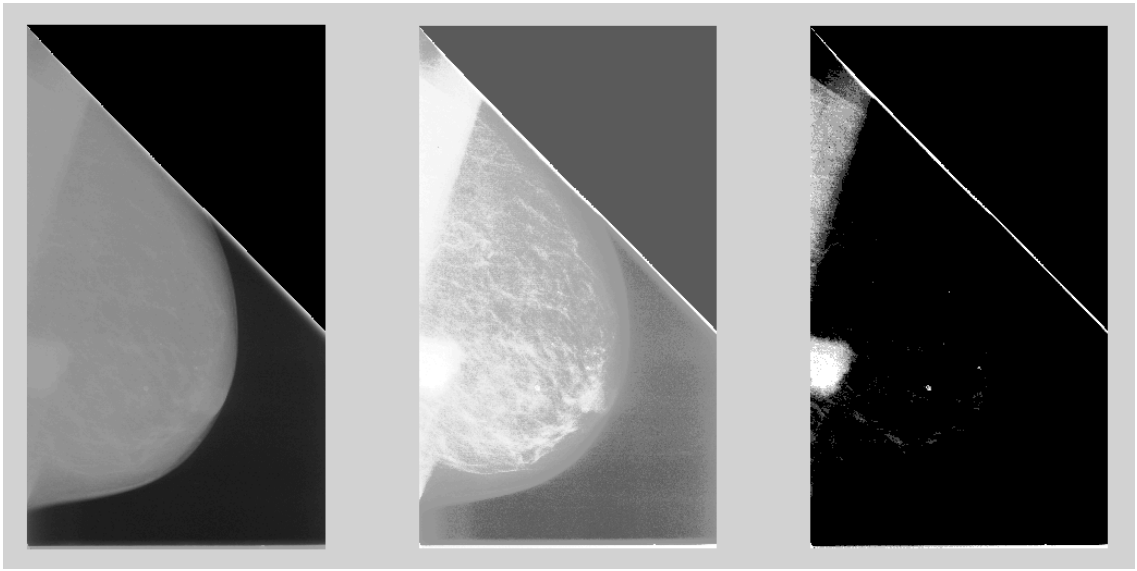


Figura 15. Procesado de nuestra imagen ejemplo A. La primera corresponde al primer histograma de la Figura 12, la segunda al primer histograma de la Figura 13 y la última al primero de la Figura 14.

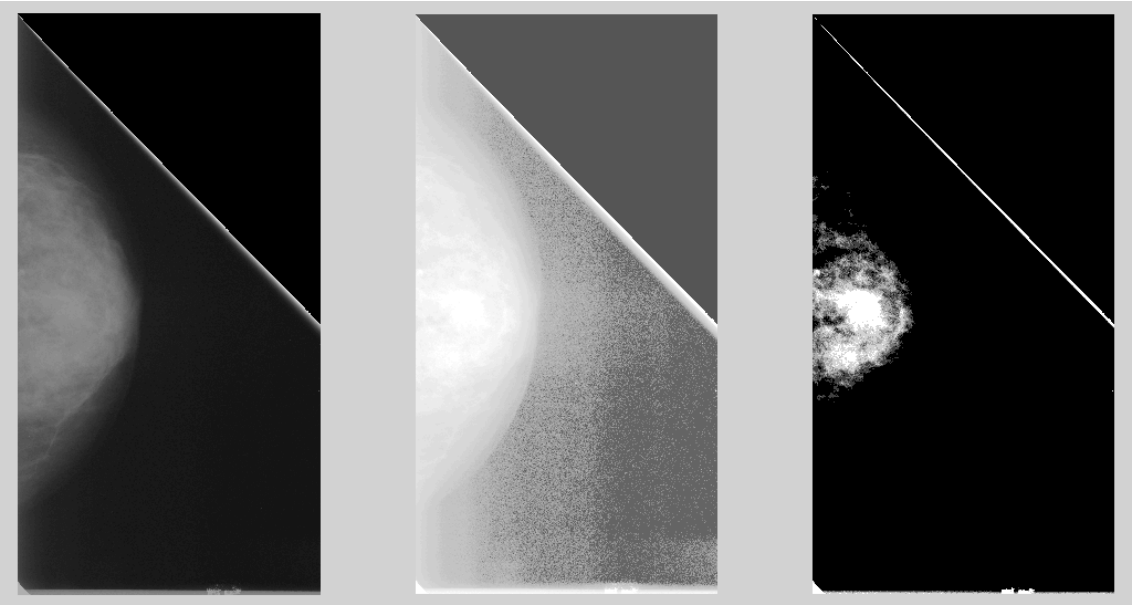


Figura 16. Procesado de nuestra imagen ejemplo A. La primera corresponde al segundo histograma de la Figura 12, la segunda al segundo histograma de la Figura 13, y la última al segundo de la Figura 14.

5.4. Justificación de la normalización

A continuación se presentan nuestras dos imágenes de estudio, pero en este caso, se les ha realizado un preprocesamiento similar al anterior, con la única diferencia de que ambas imágenes no han sido normalizadas. Se puede observar claramente la enorme diferencia que existe entre las dos imágenes, una apenas muestra la zona de la mama, mientras que la otra se muestra en su plenitud.

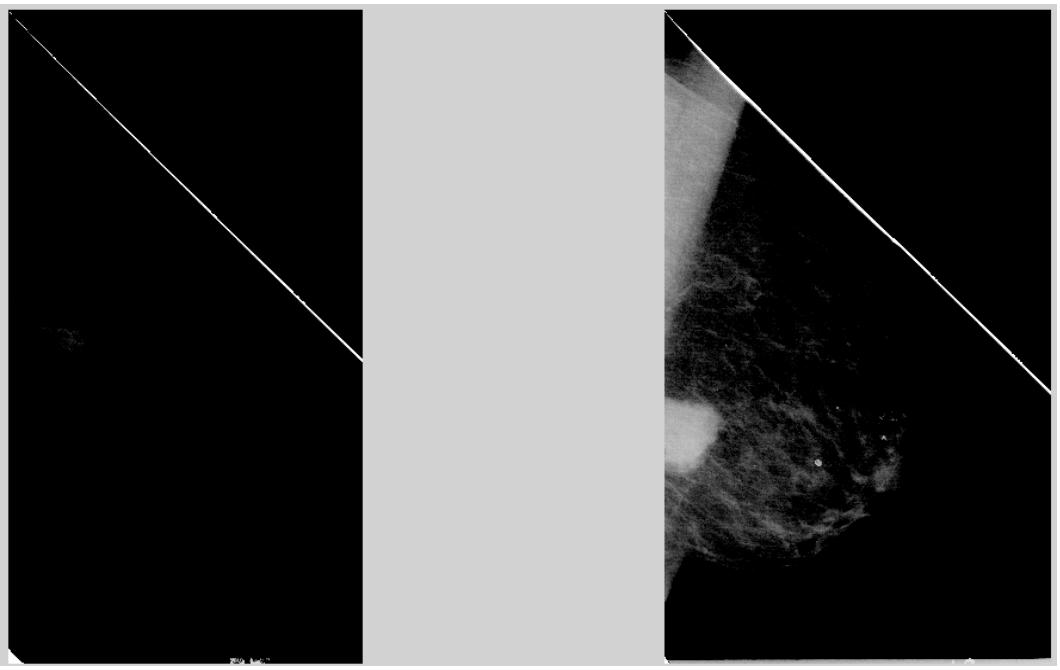


Figura 17. Preprocesado sin normalización de las imágenes ejemplo A (izquierda) y B (derecha).

6. Detección de zonas potencialmente tumorales y aislamiento de las mismas

6.1. Detección de zonas potencialmente tumorales

- Erosión con máscara en forma de disco

Para realizar la erosión, hemos utilizado una máscara con forma de disco. Dicha forma se ha indicado en MATLAB con el comando “strel” [16], introduciendo como argumento la cadena de caracteres “disk”.

Una vez creada dicha máscara, se realiza una erosión de la imagen a través del comando “imerode” [17] utilizando el elemento estructurante en forma de disco creado anteriormente. El objetivo de esta erosión es discriminar ciertas manchas de pequeño tamaño y realizar una pequeña erosión alrededor de la zona de interés. El resultado de esta erosión se encuentra indicado en la primera columna de la Figura 18.

- Erosión con máscara en forma de línea con orientación de 45º

El objetivo de esta erosión es eliminar las líneas que aún puedan quedar tras la eliminación del triángulo inicial. Para ello, hemos utilizado un procedimiento parecido al anteriormente descrito, pero utilizando una máscara en forma de línea inclinada 45º sobre el eje x para que sea perpendicular a las líneas del triángulo. El resultado de esta erosión se encuentra indicado en la segunda columna de la Figura 18.

- Discriminación de niveles de grises bajos

A partir de un valor de corte, se ha realizado un proceso de umbralización por el que todos los valores inferiores a ese valor de corte se igualan 0, de forma que los niveles de grises más altos destacan en la imagen, siendo esto uno de nuestros objetivos. El resultado de esta erosión se encuentra indicado en la tercera columna de la Figura 18.

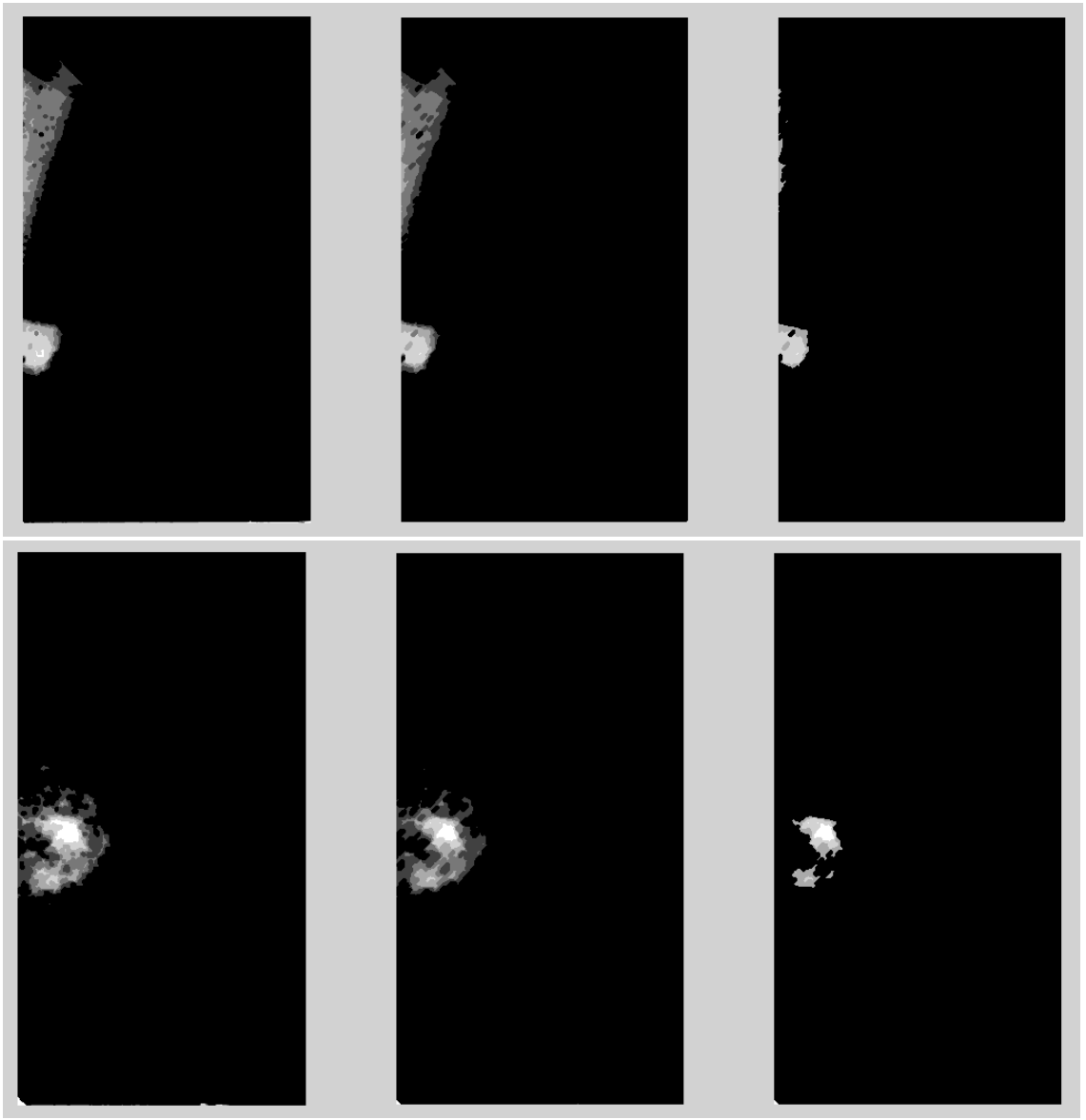


Figura 18. Segunda etapa del procesamiento de las imágenes de estudio A y B. La primera fila corresponde a la imagen A, mientras que la segunda corresponde a la imagen B.

6.2. Aislamiento de las zonas potencialmente tumorales

Tras las erosiones y la umbralización, el siguiente paso consiste en convertir la imagen a binaria. En nuestro caso, todo valor superior a 0, pasará a valer 1 (Figura 19).

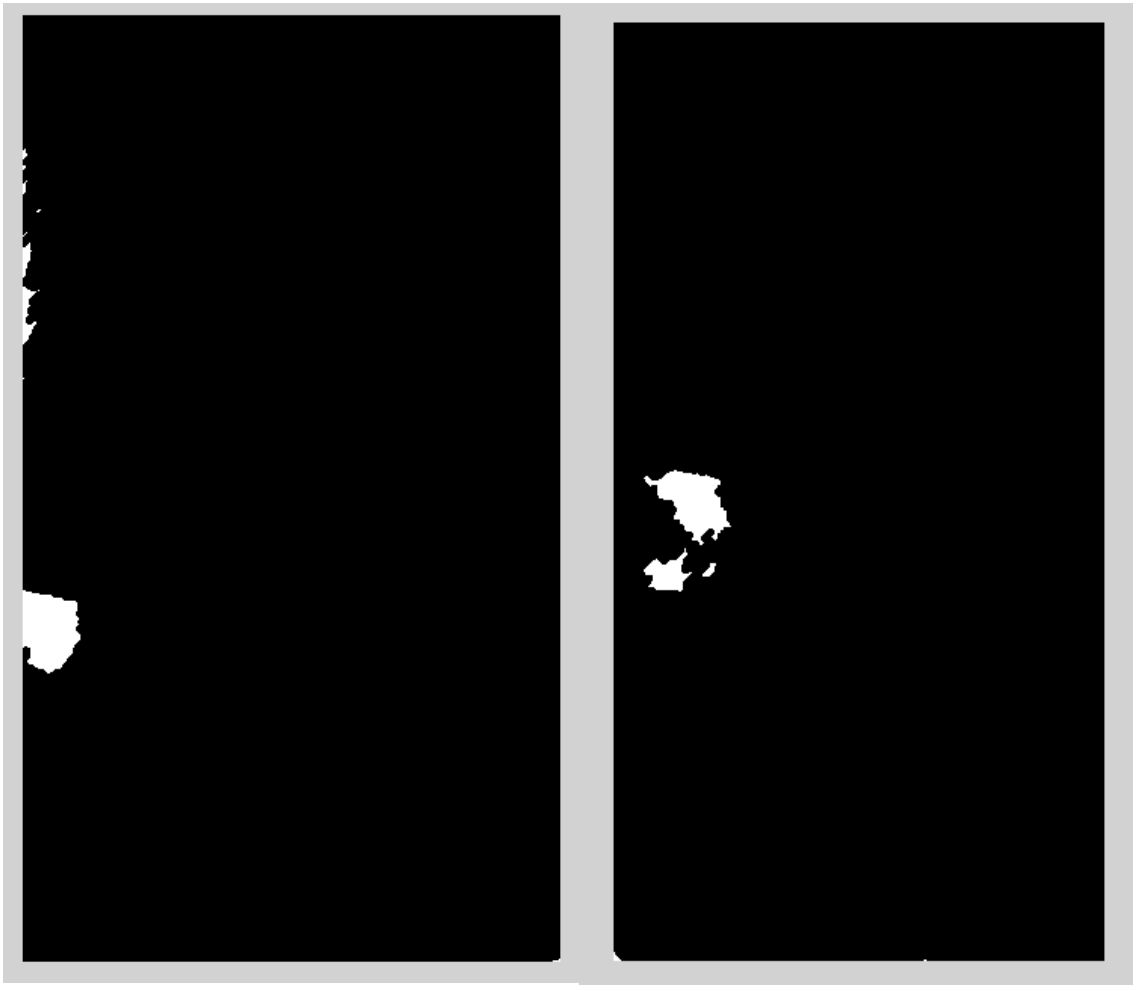


Figura 19. Resultado final del procesamiento antes del aislamiento de las manchas individuales, a la izquierda la imagen de estudio A, y a la derecha la imagen B.

Esto nos permitirá utilizar la función “regionprops” [18] para calcular los centroides de cada una de las manchas y, una vez obtenido dicho valor, utilizarlo para realizar un crecimiento de regiones en la imagen original que nos permita obtener la mancha correspondiente.

Una vez aisladas todas las manchas, se han utilizado dos criterios para decidir cual de ellas será tomada con el tumor. El primero de ellos es que el centroide calculado de la mancha se encuentre contenido en el cáncer, ya que conocemos donde se encuentra gracias al archivo OVERLAY.

En segundo lugar, y en caso de que haya varias manchas que cumplan la primera condición, el criterio que se ha utilizado para decidir cual será considerada como la mancha tumoral de la imagen, se ha empleado un algoritmo de semejanza, tomando así como tumor la mancha que sea más similar a la obtenida por el archivo OVERLAY. Nuestra imagen de estudio B es un ejemplo de este caso, y puede comprobarse en la Figura 21.

Además, hemos realizado un proceso de selección de manchas para que, a partir de unas condiciones mínimas de tamaño y localización, hemos discriminado diversas manchas. Algunas manchas presentes en la imagen que se encuentren en el extremo inferior de la imagen o tenga un tamaño muy reducido, han sido discriminadas y no se han tenido en cuenta a la hora de realizar el estudio.

El resultado final del procesamiento de nuestras dos imágenes de estudio se puede ver reflejado en las Figuras 20 y 21.

Figura 20. Resultado del aislamiento de las manchas de la imagen de estudio A. A la izquierda, la mancha obtenida por el algoritmo de detección de zonas potencialmente tumorales. A la derecha, la imagen obtenida por el algoritmo de crecimiento de regiones con la semilla indicada en el archivo OVERLAY. En este caso en concreto, solo se ha obtenido una mancha, que es la mancha tumoral.

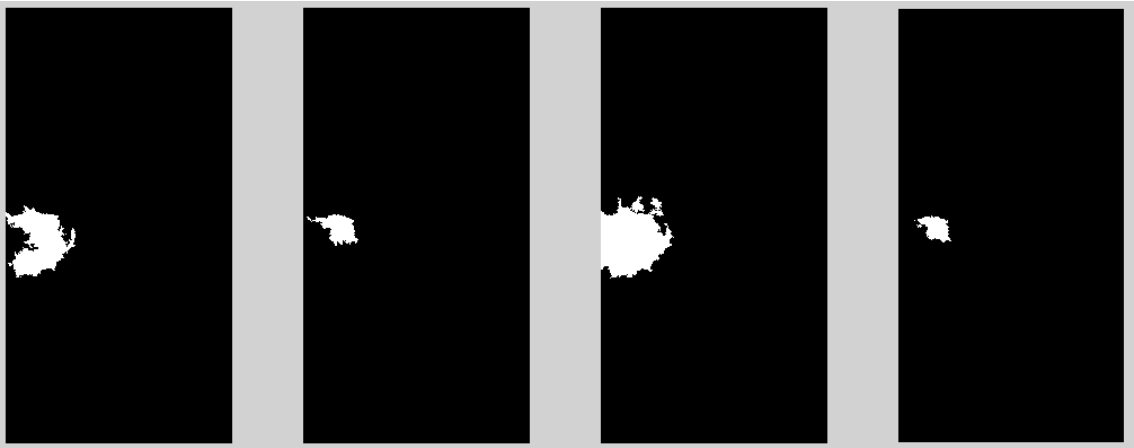


Figura 21. Resultado del aislamiento de las manchas de la imagen de estudio B. Las tres primeras imágenes corresponden a las manchas obtenidas por el algoritmo de detección de zonas potencialmente tumorales. La última es la imagen obtenida por el algoritmo de crecimiento de regiones con la semilla indicada en el archivo OVERLAY. En este caso en concreto, se han obtenido tres manchas; siendo la segunda de ellas la más similar al cáncer.

7. Estudio de los descriptores morfológicos de cada fragmento

Una región de una imagen revela diferentes tipos de características de interés debido a que, cada uno de los píxeles, almacena diferentes tonalidades de gris, describiendo un objeto. En este caso, nuestras regiones de interés son las manchas extraídas de las mamografías, obtenidas a través de un largo procesamiento de imagen. Este proceso consiste en el correcto procesamiento de la mamografía, incluyendo cambios de contraste y erosión, para la detección automática de candidatos potenciales a cáncer.

El área de procesamiento de imágenes de Matlab proporciona diversos métodos que permiten estudiar las propiedades que describen una región de la imagen; sin embargo, otras muchas características han sido calculadas a partir de algoritmos desarrollados. A continuación, se describen algunos métodos para obtener las características de intensidad, forma y textura de las manchas presentes en las mamografías [19].

7.1. Descriptores de forma

Para calcular todos estos descriptores de imagen, hemos utilizado las imágenes binarizadas de las manchas, es decir, imágenes que contienen únicamente dos valores: 0 y 1.

- Área: Se define como el número de pixeles que forman la región de interés. Para calcularla, se ha utilizado la función de MATLAB “bwarea”, con la que se calculan el número de pixeles de la imagen binarizada con valor 1.

- Perímetro: Se define como el número de pixeles que forman el contorno de la región de interés. Para calcularlo, se ha utilizado la función de MATLAB “bwperim” (Figura 22), sacando así la imagen correspondiente al perímetro de la mancha para, acto seguido, sumar el número de pixeles con valor 1 que se encuentran en esa imagen.

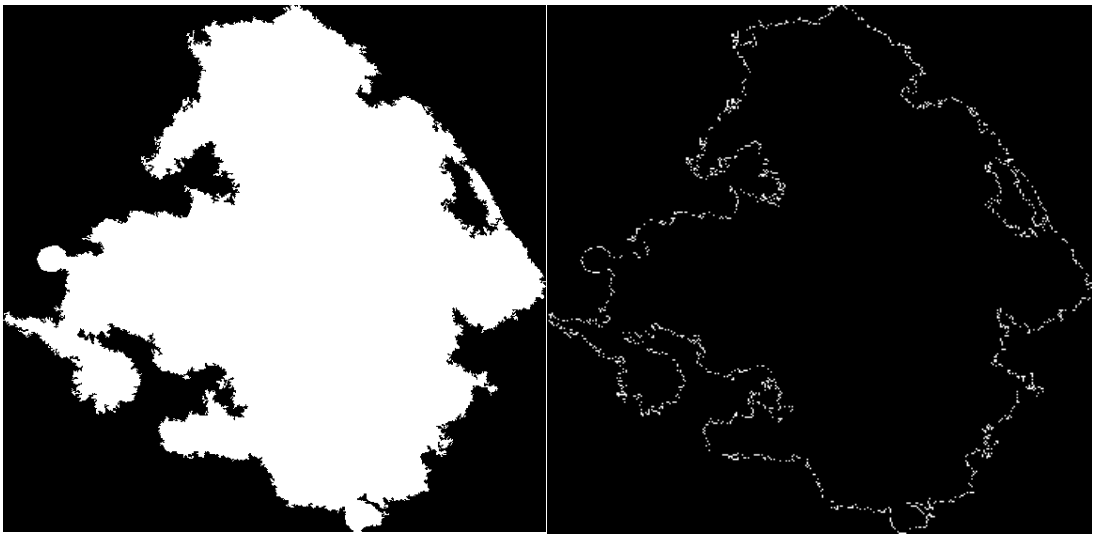


Figura 22. A la izquierda, el recorte de un cáncer en blanco y negro, y a la derecha la imagen correspondiente a su perímetro obtenida con el comando “bwperim”.

- Compacidad: Es un descriptor que define cuan compacto es el objeto de interés. Para calcularlo, hemos aplicado la siguiente fórmula [20]:

$$Compacidad = \frac{A}{P^2}$$

Siendo A el área de la región y P el perímetro de esa región.

- Circularidad: Es un descriptor que define cuan circular es el objeto de interés. Para calcularlo, hemos aplicado la siguiente fórmula [10]:

$$Circularidad = 4\pi \frac{A}{P^2}$$

Siendo A el área de la región y P el perímetro de esa región.

- Asimetría: Este descriptor nos permite identificar si los pixeles de la zona de interés se distribuyen de forma uniforme alrededor del punto central, representado por la media aritmética. Para calcular la asimetría, tendremos que calcular el momento centrado de orden 3 con la siguiente fórmula:

$$\mu_k = E[(X - E[X])^k] = \int_{-\infty}^{+\infty} (x - \mu)^k f(x)\, dx$$

Siendo μ el momento a calcular, E el operador correspondiente a la esperanza, k el orden y X una variable aleatoria. Calculando el momento centrado de orden 3, tenemos lo siguiente:

$$E[(x - m)^3] = \mu_3 - 3\mu\sigma^2 - \mu^3$$

Para calcularlo de forma sencilla, se puede utilizar el comando “moment” de MATLAB [21].

- Curtosis: Este descriptor determina el grado de concentración que presentan los pixeles en la región central de la distribución, representada por la media aritmética. Para calcular la curtosis, tendremos que calcular el momento centrado de orden 4 con la siguiente fórmula:

$$\mu_k = E[(X - E[X])^k] = \int_{-\infty}^{+\infty} (x - \mu)^k f(x)\, dx$$

Siendo μ el momento a calcular, E el operador correspondiente a la esperanza, k el orden y X una variable aleatoria. Calculando el momento centrado de orden 3, tenemos lo siguiente:

$$E[(x - m)^4] = \mu_4 - 4\mu\mu_3 + 6\mu^2\sigma^2 - 3\mu^4$$

Para calcularlo de forma sencilla, se puede utilizar el comando “kurtosis” de MATLAB [22].

7.1.1. Longitud radial normalizada (LRN)

La Longitud radial normalizada es definida como la distancia Euclidiana normalizada, e(i, j), entre el centroide de la zona de interés hacia uno de los puntos de su contorno. La distancia se normaliza al considerar la máxima distancia euclidiana que existe hacia un punto del contorno. Para calcular la distancia euclidiana, se utiliza la siguiente fórmula:

$$DistanciaEuclidiana = \sqrt{(x_{centroide} - x)^2 + (y_{centroide} - y)^2}$$

Ahora, debemos proceder a normalizarla:

$$LRN_{normalizado} = \frac{LRN}{|LRN|}$$

Los descriptores de forma que podemos extraer a partir de los valores de la longitud radial normalizada son los siguientes:

- Media: Describe el promedio de la distancia existente entre el centroide de la región y un punto del contorno de la región de interés. Se calcula a través de la siguiente fórmula:

$$e_{media} = \frac{1}{P}\sum_{i=1}^{p} e(i)$$

- Desviación típica: Este descriptor estima las irregularidades del contorno de la región de interés. El valor de este descriptor es mayor cuando se presentan más irregularidades en el contorno. Se calcula a través de la siguiente fórmula:

$$\sigma = \sqrt{\frac{1}{P}\sum_{i=1}^{p}(e(i) - e_{media})^2}$$

- Entropía: Es un descriptor que se puede obtener a partir del histograma de la longitud radial normalizada (LRN). Para calcular este valor, hemos utilizado el comando de MATLAB “wentropy” [23].

- Rugosidad del contorno: Se utiliza para describir las irregularidades del contorno en relación a los picos y salientes que puede presentar la región de interés. Se calcula a través de la siguiente fórmula:

$$RC = \frac{1}{P}\sum_{i=1}^{p}|e(i) - e(i+1)|$$

- Cruce por cero (CPC): Descriptor que se calcula contando el número de veces en que el valor de la longitud radial normalizada (LRN) supera a su media. Se calcula a través de la siguiente fórmula:

$$CPC = \sum_{i=1}^{p} h(e(i), e_{media})$$

Cuando el valor de $e(i)$ sea mayor que el valor de e_{media}, la función $h(e(i), e_{media})$ valdrá 1. En cualquier otro caso, dicha función valdrá 0.

7.2. Descriptores de intensidad

Este tipo de descriptores utilizan como punto de partida la mancha objetivo en niveles de grises y se basan en estudiar como se comportan estos en toda la sección de la mancha. Tomaremos en adelante I(x , y) como el valor de intensidad del pixel de coordenadas (x , y).

- Asimetría. Representa la distribución de las intensidades de los píxeles hacia los extremos con respecto al punto central [14].

$$Asimetría = \frac{1}{NM\sigma^2} \sum_{x=1}^{N} \sum_{y=1}^{M} (I(x,y) - \mu)^3$$

- Curtosis. Medida de forma que representa el grado en que la distribución de intensidades está escarpada o achatada [24] (Figura 23).

$$Curtosis = \frac{1}{NM\sigma^4} \sum_{x=1}^{N} \sum_{y=1}^{M} (I(x,y) - \mu)^4 - 3$$

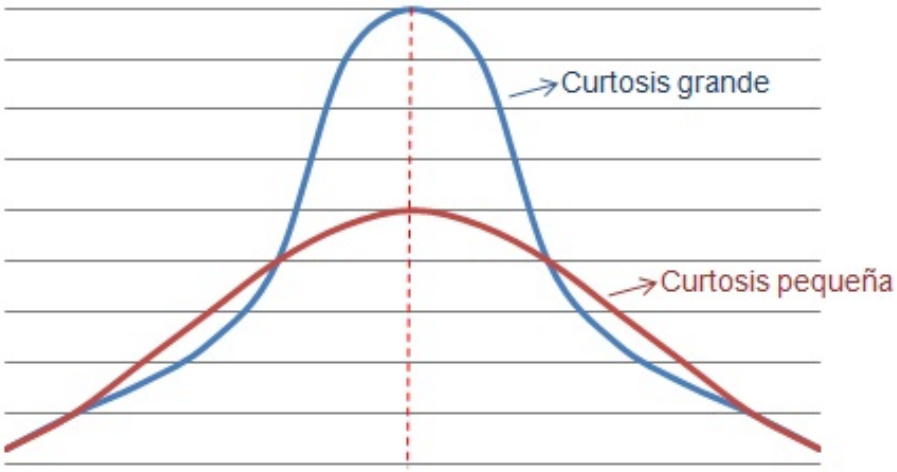


Figura 23. Diferencia entre valores de curtosis altos o bajos.

- Media. Representa el valor promedio de intensidad de la mancha.

$$\mu = \frac{1}{NM} \sum_{x=1}^{N} \sum_{y=1}^{M} I(x,y)$$

- Mediana. Representa el valor de intensidad de la posición central en el conjunto de las intensidades ordenadas.

- Varianza. Representa el grado de dispersión de los valores de intensidad con respecto a la media.

$$\sigma^2 = \frac{1}{NM} \sum_{x=1}^{N} \sum_{y=1}^{M} (I(x,y) - \mu)^2$$

7.3. Textura lineal

Las masas malignas suelen tener un aspecto mucho más irregular que las manchas normales. Estas manchas malignas, en muchos de los casos, se encuentran rodeadas de una radiación de líneas en forma de pico. Muchas veces la densidad es muy débil y cuando se encuentra incrustado en el tejido normal de la mama puede ser muy difícil de percibir.

Dada la magnitud $M(x,y)$ y la fase $\varphi(x,y)$ se construye la suma de los vectores de doble ángulo, representada a través de la siguiente ecuación:

$$\mathrm{z} = \mathrm{C}e^{2\theta} = C * \cos(2\theta) + C * \sin(2\theta)$$

Donde $C = M(x,y)$ y $\theta = \varphi(x,y)$.

Para obtener el descriptor de textura lineal, se calcula la longitud de los componentes $M(x,y)$ y $\varphi(x,y)$ mediante la siguiente ecuación:

$$z_1 = \sqrt{\left(\sum \mathrm{C} * \cos(2\theta)\right)^2 + \left(\sum \mathrm{C} * \sin(2\theta)\right)^2}$$

Acto seguido, procedemos a calcular la longitud total de todos los vectores con la siguiente fórmula:

$$z_2 = \sum \sqrt{(C * \cos(2\theta))^2 + (C * \sin(2\theta))^2}$$

Con estas longitudes calculadas, pasamos a calcular el valor del descriptor de textura lineal, que sería el siguiente:

$$TexturaLineal = \frac{z_1}{z_2}$$

7.4. Matriz de co-ocurrencia de niveles de grises (GLCM)

Una matriz de co-ocurrencia es una matriz que se define sobre una imagen e indica la distribución de los valores concurrentes a un determinado desplazamiento. Matemáticamente, una matriz de co-ocurrencia C es definida sobre una imagen Im representada por una matriz de tamaño NxM, parametrizado por un desplazamiento dado por $(\Delta x, \Delta y)$ de la siguiente manera [25]:

$$C_{\Delta x,\Delta y}(i,j) = \sum_{p=1}^{n}\sum_{q=1}^{m}\begin{cases}1, si\ Im(p,q) = i\ y\ Im(p+\Delta x, q+\Delta y) = j\\ 0, de\ cualquier\ otra\ manera\end{cases}$$

Donde i y j representan el valor de la intensidad de la imagen, p y q son las posiciones de la imagen Im y el desplazamiento $(\Delta x, \Delta y)$ depende del ángulo θ y la distancia d a la cual la matriz es computarizada.

Los descriptores que se pueden extraer de la matriz de co-ocurrencia de niveles de grises son los siguientes:

- Contraste: Este descriptor permite medir variaciones fuertes o bruscas de los niveles de intensidad de la imagen. Este descriptor se calcula a partir de la siguiente fórmula:

$$Contraste = \sum_{n=0}^{Ng-1} i^2 \left\{\sum_{i=1}^{Ng}\sum_{j=1}^{Ng} C(i,j), |i-j| = n\right\}$$

- Energía: Mide el grado de homogeneidad de una imagen. Este descriptor se calcula a partir de la siguiente fórmula:

$$Energia = \sqrt{\sum_{i=1}^{Ng}\sum_{j=1}^{Ng}\big(C(i,j)\big)^2}$$

- Correlación: Mide la dependencia lineal de los niveles de gris entre los píxeles y posiciones específicas relacionadas con cada uno de ellos dentro de la matriz. Se calcula a través de la siguiente fórmula:

$$Correlacion = \sum_{i=1}^{Ng}\sum_{j=1}^{Ng}\frac{(i-\mu_x)\big(i-\mu_y\big)C(i,j)}{\sigma_x\sigma_y}$$

Donde μ_x, μ_y, σ_x y σ_y son las medias y desviaciones típicas de C_x y C_y. Cada uno de estos valores se obtiene de la siguiente manera:

$$\mu_x = \sum_{i=1}^{Ng} iC_x \quad y \quad \mu_y = \sum_{j=1}^{Ng} jC_y$$

$$\sigma_x = \sum_{i=1}^{Ng} (i-\mu_x)^2 C_x \quad y \quad \sigma_y = \sum_{j=1}^{Ng} (j-\mu_j)^2 C_y$$

$$C_x(i) = \sum_{i=1}^{Ng} C(i,j) \quad y \quad C_y(y) = \sum_{j=1}^{Ng} C(i,j)$$

- Media: Describe el promedio de nivel de grises de la región de interés. Se calcula a través de la siguiente fórmula:

$$Media = \frac{1}{P} \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} C(i,j)$$

- Varianza: Describe la definición y agrupación de los elementos de la matriz. Este valor aumenta cuando la distancia entre elementos de la matriz con respecto a la diagonal principal es baja. Se calcula a través de la siguiente fórmula:

$$Varianza = \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} (i-\mu)^2 C(i,j)$$

- Homogeneidad: Es un valor que depende de la diagonal principal de la matriz. Este valor será grande si los valores de la diagonal principal de la matriz son grandes. Se calcula a través de la siguiente fórmula:

$$Homogeneidad = \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} \frac{1}{1+(i-j)} C(i,j)$$

Hemos utilizado las funciones “graycomatrix” [26] y “graycoprops” [27] para calcular estos valores en Matlab.

7.5. Matriz de longitudes de secuencias de niveles de grises (GLRLM)

Una secuencia de nivel de gris es un conjunto consecutivo de pixeles con el mismo nivel de gris y en una dirección dada. La longitud de dichas secuencias es el número de píxeles que contienen; y el número de veces que una secuencia aparece en una imagen es el valor de la longitud de secuencia.

La matriz de longitudes de secuencia de niveles de grises (GLRLM) es una matriz bidimensional donde cada posición P(i , j | θ) indica el número de veces que aparece una secuencia de longitud j con nivel de gris i en una dirección θ, que en nuestro estudio θ siempre valdrá 0 [28].

El número de niveles de grises de una imagen se suele disminuir realizando una recuantización, normalmente en 16 niveles de grises, lo cual es suficiente para el análisis de las texturas. En nuestro caso el algoritmo que procesa la matriz realiza este proceso, además de crear una máscara que indica la región de interés, la cual puede ser creada de tal forma que indique que dicha área de interés es la propia imagen completa.

Una vez se calcula la matriz GLRLM, se pueden extraer de ella diversas características de interés que explicaremos a continuación [29] [30]. Tomaremos G como el número de niveles de grises y L como la secuencia más larga.

- Énfasis de secuencia corta (Short Run Emphasis). Este valor aumenta cuando las secuencias que dominan son las cortas.

$$SRE = \sum_{0}^{G-1} \sum_{1}^{L} \frac{P(g,l)}{l^2}$$

- Énfasis de secuencia larga (Long Run Emphasis). Este valor aumenta cuando las secuencias que dominan son las largas.

$$LRE = \sum_{0}^{G-1} \sum_{1}^{L} P(g,l)\, l^2$$

- No uniformidad de los niveles de grises (Gray Level Non-Uniformity). Este valor aumenta cuando valores de grises aislados dominan el histograma.

$$GLNU = \sum_{1}^{L} \left[\sum_{0}^{G-1} P(g,l) \right]^2$$

- No uniformidad de la longitud de secuencia (Run Length Non-Uniformity). Este valor aumenta cuando pocos valores de longitud de secuencia aislados dominan el histograma.

$$RLNU = \sum_{0}^{G-1} \left[\sum_{1}^{L} P(g,l) \right]^2$$

- Énfasis de secuencia de grises baja (Low Gray level Run Emphasis). El énfasis es ortogonal a SRE, aumentando los valores cuando en la textura dominan muchas secuencias de valor de gris bajo.

$$LGRE = \sum_{0}^{G-1} \sum_{1}^{L} \frac{P(g,l)}{(g+1)^2}$$

- Énfasis de secuencia de grises alta (High Gray level Run Emphasis). El énfasis es ortogonal a LRE, aumentando los valores cuando en la textura dominan muchas secuencias de valor de gris alto.

$$HGRE = \sum_{0}^{G-1} \sum_{1}^{L} P(g,l)\,(g+1)^2$$

- Porcentaje de secuencia (Run Percentage). Este valor da información de la homogeneidad general del histograma, y su valor es máximo cuando todas las secuencias son de longitud unidad sin tener en cuenta su nivel de gris.

$$RP = \sum_{0}^{G-1} \sum_{1}^{L} \frac{1}{P(g,l)\,l}$$

Estos valores han sido calculados con la función “glrlm” [31] en Matlab.

8. Creación de la matriz de datos para la red neuronal

El último paso que compete a este trabajo es la creación de una matriz que servirá de base al trabajo paralelo para la clasificación de los datos por parte de la red neuronal. Esta matriz indicará los resultados normalizados obtenidos de cada uno de los descriptores para todas las manchas, indicando además la clase de cada mancha, esto es, si es una mancha cancerígena o no. Por tanto, la matriz presentará el siguiente formato:

- Cada fila indica cada una de las manchas
- Cada columna indica cada uno de los descriptores
- La última columna, sin embargo, indica la clase (1 para indicar que presentan un tumor, y 0 para indicar que no lo presentan)

9. Resultados

En esta sección del trabajo, se presentarán dos apartados de resultados diferentes. El primero de ellos está relacionado con el presente trabajo, otorgando una visión global del resultado del procesamiento de las imágenes, y analizando los costes de computación. El segundo apartado versará sobre el trabajo realizado por Rodrigo Culotta López, esto es, los resultados obtenidos de la clasificación de las manchas aisladas por la red neuronal, cerrando así el círculo y dando una idea general del resultado de todo este proceso.

9.1. Resultados del presente trabajo

En esta sección analizaremos los resultados del procesamiento de las 138 imágenes con las que se ha trabajado en este proyecto, 51 imágenes que presentan algún tipo de tumoración; y 87 imágenes sin tumoración ninguna.

Todo el procesamiento de las imágenes explicado en apartados anteriores, ha dado lugar a un conjunto de 628 zonas potencialmente tumorales de estudio. Esto supone que por cada imagen que se analiza, se extraen una media de 4.55 zonas de estudio.

Este último valor es el que nos permite ver el resultado real del procesamiento, ya que cuantas menos zonas deriven de cada imagen, menor será el volumen de análisis posterior. Cualquier mejora en el procesamiento actual que permita eliminar zonas irrelevantes, ya sea por la posición en la que se encuentren, su tamaño, su intensidad…, derivará en una reducción de este valor, y por tanto, en un resultado mejor.

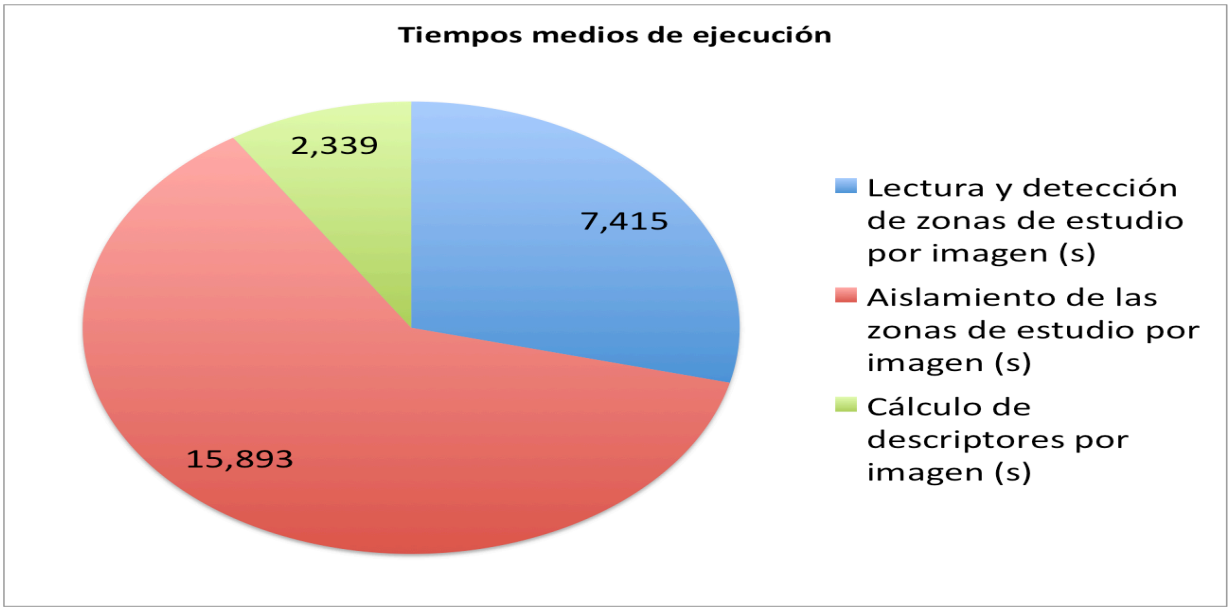
El principal problema que se puede analizar en los resultados del procesamiento es la detección de ciertas zonas fuera del área de la mama que son tomadas como relevantes. Esto es debido a que en ciertas imágenes están presentes letras que se utilizan para dar información del tipo de mamografía y de la mama que se está escaneando; o bien hay zonas ruidosas intensas que también son aisladas.

Para solucionar estos problemas, se debería haber utilizado algún tipo de algoritmo que detectara y delimitara la zona de la mama; y así excluir la zona restante. En la Figura 24, se presentan dos ejemplos de las letras antes mencionadas, que corresponden a dos de las zonas de estudio aisladas.

Figura 24. Ejemplos de letras aisladas tras el procesamiento.

A continuación, analizaremos los tiempos de ejecución empleados en el procesamiento de cada una de las partes en las que se encuentra dividido el trabajo; dándonos una idea de cuáles son las tareas más costosas de todo el proceso.

Tiempos medios de ejecución	***Segundos***
Lectura y detección de zonas de estudio por imagen (s)	7,415
Aislamiento de las zonas de estudio por imagen (s)	15,893
Cálculo de descriptores por zona de estudio (s)	0,514
Cálculo de descriptores por imagen (s)	2,339
Total empleado por imagen (s)	25,647



Cabe destacar el hecho de que el tiempo indicado como “Cálculo de descriptores por zona de estudio”, corresponde al tiempo en calcular los 11 descriptores definitivos; tras realizar una selección de los más importantes para la clasificación de la red neuronal. Sin embargo, no es mucha la diferencia entre este tiempo y el tiempo que se empleaba en calcular los 32 descriptores iniciales con los que se comenzó el estudio; mientras que para 11 descriptores se emplean unos 0,514 segundos, para calcular 32 descriptores se necesitaban 0,628 segundos. Esto nos permite darnos cuenta de que el cálculo de los descriptores es claramente la tarea más sencilla de llevar a cabo, y que la cantidad de descriptores no influye demasiado.

9.2. Resultados del trabajo paralelo

Estos resultados son la consecuencia de todo el proceso del proyecto, son fruto de analizar y clasificar toda la información que deriva de este trabajo, y dan por concluido el proyecto. Se basan en los resultados obtenidos por la red neuronal después de procesar las imagen que han derivado del presente trabajo.

Es importante recordar que nuestro principal objetivo es evitar obtener falsos negativos, lo cual supondría haber clasificado correctamente todas las zonas cancerígenas. El segundo objetivo será reducir al máximo el número de falsos positivos, procurando obtener un porcentaje alto de aciertos.

A continuación, se expondrá un breve resumen de estos resultados, suficiente para dar una idea del rendimiento obtenido y centrándonos en el porcentaje total de acierto y en el número de falsos negativos obtenidos, sin entrar en mucho detalle; ya que para ello puede consultarse el trabajo paralelo *“Detección y clasificación de tumores en mamografías a través de redes neuronales”.*

Diferenciaremos tres resultados obtenidos y considerados como definitivos:

- Utilizando el software de clasificación “WEKA”, con un clasificador Naive Bayes, se ha obtenido un porcentaje de acierto del 85.669 % y se han clasificado erróneamente 2 cánceres.

- Utilizando una red neuronal Fitnet de Matlab con función de entrenamiento bayesiana, se ha logrado obtener un 98.885 % de acierto, pero se han obtenido 5 falsos negativos.

Por último, realizando un estudio más profundo de los resultados, se ha detectado que las clasificaciones realizadas por redes neuronales Fitnet con función de entrenamiento Levenberg-Marquardt eran excelentes (Figura 25).

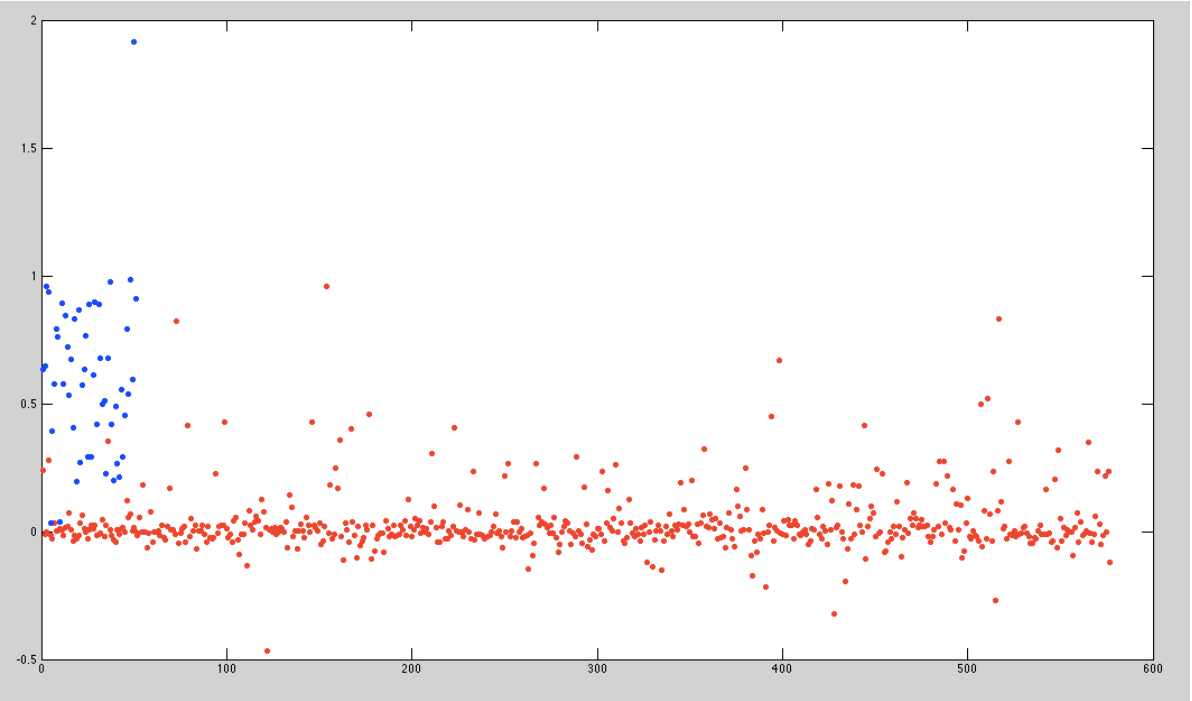


Figura 25. Ejemplo de clasificación de una red neuronal Fitnet con función de entrenamiento Levenberg-Marquardt. En azul, las manchas cancerígenas; en rojo, las manchas normales.

El hecho de que se observe una notable diferencia entre las manchas con cáncer y las manchas normales, motivó a hacer un nuevo estudio con valores de corte superiores, obteniendo unos mejores resultados.

En concreto, se ha conseguido obtener un porcentaje de acierto del 90.127 %, mientras que la clasificación correcta de las manchas cancerígenas ha sido del 100%; lo cual implica haber cumplido el principal objetivo de este proyecto.

Respecto a los costes computacionales de estas clasificaciones, pueden ser considerados como muy bajos respecto al volumen de datos con el que trabaja, menos de 10 segundos en cualquier tipo de clasificación; lo cual puede considerarse un resultado casi instantáneo para un software médico.

Cabe destacar que las redes neuronales, ni han sido reentrenadas, ni han sido entrenadas con todo el volumen de manchas (un 15% de las mismas solo eran utilizadas en la clasificación). Esto asegura unos resultados más reales ya que ese 15% es clasificado sin ser previamente estudiado por la red neuronal.

10. Conclusiones

- Los resultados del procesamiento de las imágenes para la detección de los cánceres potenciales han sido muy satisfactorios. No se ha pasado por alto ningún cáncer, se han discriminado muchas manchas irrelevantes, y no se han obtenido un gran número de manchas de cada imagen.

- Aunque el procesamiento de las imágenes haya sido bueno, se han obtenido manchas correspondientes a letras o artefactos extraños presentes en las mamografías. Esto se podría mejorar realizando en las primeras fases del proceso un aislamiento de la zona de la mama, lo que reduciría el número total de manchas localizadas.

- Los resultados obtenidos en la clasificación de las imágenes han sido buenos, obteniendo un porcentaje alto de aciertos, concretamente un 90%; y clasificando correctamente todas las zonas cancerígenas. Estos resultados mejorarían de poder desarrollarse este método con unos medios más homogéneos y un banco de imágenes más completo.

- Finalizado este software, sería ideal poder realizar un trabajo de validación en un hospital con imágenes de sus bases de datos; además de realizar una interfaz que permita visualizar las imágenes y facilite el uso del programa. En concreto, estudiar las mamografías tomadas por un tipo de equipo particular sería lo óptimo, para adaptar el software, si fuese necesario, a los parámetros con los que son tomadas las imágenes por los aparatos.

- Estamos convencidos de que trabajando más en este tema, con mejores medios y más tiempo, se podrían conseguir unos resultados excelentes y se podría desarrollar un programa que podría incluso tener repercusión en la práctica clínica.

Bibliografía

[1] Michie, D., Spiegelhalter, D. J., & Tayor, C. C. (1994). Machine learning, neural and statistical classification. London: Ellis Horwood.

[2] Aragonés, M. J., Ruiz, A. G., Jiménez, R., Pérez, M., & Conejo, E. A. (2003). A combined neural network and decision trees model for prognosis of breast cancer relapse. Artificial Intelligence in Medicine, 27, 45–63.

[3] Choua, S.-M., Leeb, T.-S., Shaoc, Y. E., & Chenb, I.-F. (2004). Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. Expert Systems with Applications, 27, 133–142.

[4] Ryua, Y. U., Chandrasekaranb, R., & Jacobc, V. S. (2007). Breast cancer prediction using the isotonic separation technique. European Journal of Operational Research, 181, 842–854.

[5] Sahan, S., Polat, K., Kodaz, H., & Günes, S. (2007). A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. Computers in Biology and Medicine, 37, 415–423.

[6] Übeyli, E. D. (2007). Implementing automated diagnostic systems for breast cancer detection. Expert Systems with Applications, 33, 1054–1062.

[7] Fujifilm.
<http://www.fujifilm.eu/es/productos/sistemas-medicos/mamografia-digital/>
[Consulta: Septiembre de 2015].

[8] General Electric.
<http://www3.gehealthcare.es/es-es/productos/categorias/mamografia#tabs/tab4A12959A46654E59A0C36C5746FF31C6>
[Consulta: Septiembre de 2015].

[9] Siemens. <http://www.healthcare.siemens.es/mammography>
[Consulta: Septiembre de 2015].

[10] Philips.
<http://www.philips.es/healthcare/solutions/mammography/digital-mammography>
[Consulta: Septiembre de 2015].

[11] University of South Florida. Digital Mammography Home Page.
DDSM: Digital Database for Screening Mammography.
<http://marathon.csee.usf.edu/Mammography/Database.html>
[Consulta: Septiembre de 2015].

[12] DDSM Software. Dr. Chris Rose, University of Manchester. <http://microserf.org.uk/academic/Software.html> [Consulta: Septiembre de 2015].

[13] Regiongrow function.
<fourier.eng.hmc.edu/e161/dipum/regiongrow.m> [Consulta: Septiembre de 2015].

[14] Función “imadjust”.
<http://es.mathworks.com/help/images/ref/imadjust.html> [Consulta: Septiembre de 2015].

[15] Función “histeq”.
<http://es.mathworks.com/help/images/ref/histeq.html> [Consulta: Septiembre de 2015].

[16] Función “strel”.
<http://es.mathworks.com/help/images/ref/strel.html> [Consulta: Septiembre de 2015].

[17] Función “imerode”.
<http://es.mathworks.com/help/images/ref/imerode.html> [Consulta: Septiembre de 2015].

[18] Función “regionprops”.
<http://es.mathworks.com/help/images/ref/regionprops.html> [Consulta: Septiembre de 2015].

[19] Pedro Gabriel, Rolando (2014). Tesis doctoral “Clasificación de masas en imágenes de mamografías utilizando redes bayesianas” p. 21-38 <http://jupiter.utm.mx/~tesis_dig/12183.pdf> [Consulta: Septiembre de 2015].

[20] Dr. José Muñoz Pérez. Apuntes de Representación de Formas. Departamento de Lenguajes y Ciencias de la Computación. Universidad de Málaga.
<http://www.lcc.uma.es/~munozp/documentos/procesamiento_de_imagenes/temas/Pres-Tema8%20PI.ppt> [Consulta: Septiembre de 2015].

[21] Función “moment”.
<http://es.mathworks.com/help/stats/moment.html> [Consulta: Septiembre de 2015].

[22] Función “kurtosis”.
<http://es.mathworks.com/help/stats/kurtosis.html> [Consulta: Septiembre de 2015].

[23] Función “wentropy”.
<http://es.mathworks.com/help/wavelet/ref/wentropy.html> [Consulta: Septiembre de 2015].

[24] Asimetría y Curtosis. Web Universo fórmulas.
<http://www.universoformulas.com/estadistica/descriptiva/asimetria-curtosis/>
[Consulta: Septiembre de 2015].

[25] M. Haralick, Robert; Shanmugam, K.; Dinstein, Its’hak (1973). Textural Features for Image Classification.
<http://haralick.org/journals/TexturalFeatures.pdf> [Consulta: Septiembre de 2015].

[26] Función “graycomatrix”.
<http://es.mathworks.com/help/images/ref/graycomatrix.html> [Consulta: Septiembre de 2015].

[27] Función “graycoprops”.
<http://es.mathworks.com/help/images/ref/graycoprops.html> [Consulta: Septiembre de 2015].

[28] Albregtsen, Fritz (1995). “Statistical Texture Measures Computed from Gray Level Run Length Matrices”. University of Oslo.
<http://heim.ifi.uio.no/~in384/info/glrlm.ps> [Consulta: Septiembre de 2015].

[29] Tang, Xiaoou (1998). Texture Information in Run-Length Matrices.
<http://www.res<earchgate.net/publication/3326868_Texture_information_in_run-length_matrices> [Consulta: Septiembre de 2015].

[30] Haidekker, Mark (2010). Advanced Biomedical Image Analysis p. 260.

[31] Gray level length image statistics function. Mathworks.
<http://www.mathworks.com/matlabcentral/fileexchange/52640-gray-level-run-length-image-statistics/content/glrlm.m> [Consulta: Septiembre de 2015].